DOCUMENT RESUME

ED 193 240 TM 800 554

TITLE	American Statistical Association: 1979 Proceedings of the Section on Statistical Education (22nd, Washington, D.C., August 13-16, 1979).
INSTITUTION	American Statistical Association, Washington, D.C.
PUB DATE	13 Aug 79
NOTE	100p.: Small print may be marginally legible.
AVAILABLE FROM	American Statistical Association, 806 15th St., N.W., Washington, DC 20005 ($6.00).
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Higher Education; Mathematical Models; *Mathematics Education; *Mathematics Instruction; Physicians; *Professional Education; Research Design; Statistical Analysis; *Statistics; *Student Evaluation of Teacher Performance; Teaching Methods; Testing

ABSTRACT

Twenty-two papers and discussions presented at the 1979 annual meeting of the American Statistical Association are reproduced. Papers dealing with student evaluation of instruction include "An Analysis of Student Evaluation of Instruction and Relationships Among Salary, Chairman's Ratings, and Student Ratings of Faculty," by Alvin C. Rencher, "Administrative Use of Student Surveys: One View," by Paul D. Minton; and "Students' Evaluations: Their Antecedents and Aftermath," by W. F. Seibert and A.R. Starry. A "Report of the Committee on Training of Statisticians for Industry," is included, as well as remarks amplifying the report by Paul D. Minton. The topic of teaching service courses on statistics is covered by the following papers: "A Modular Self-Paced Statistics Course: Five Years' Experience," by Donald A. Anderson and William P. Cocke; "Student's Choice: Self-Pacing and/or the Lecture Method for Introductory Statistics," by Howard B. Christensen; and "The Elementary Statistics Course: Fact or Fiction," by J. Leroy Folks. "Teaching Statistics to Medical Specialists: Cardiologists," by James H. Ware, and a discussion following it by Dennis B. Gillings complete the theme papers. Thirteen additional papers are provided. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED193240

$6.00

AMERICAN STATISTICAL ASSOCIATION

1979 PROCEEDINGS

of the

SECTION ON STATISTICAL EDUCATION

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Papers presented at the Annual Meeting of the American Statistical Association, Washington, D.C., August 13-16, 1979 under the sponsorship of the Section on Statistical Education

Price: $6.00

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
American Statistical Assn
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 800554

AMERICAN STATISTICAL ASSOCIATION
806 15th Street, N.W. Washington, D.C. 20005

AMERICAN STATISTICAL ASSOCIATION

1979
PROCEEDINGS

of the

SECTION ON STATISTICAL EDUCATION

Papers presented at the Annual Meeting of the American Statistical Association, Washington, D.C., August 13-16, 1979 under the sponsorship of the Section on Statistical Education

AMERICAN STATISTICAL ASSOCIATION
806 15th Street, N.W. Washington, D.C. 20005

OCT 3

INDEX OF PARTICIPANTS

Anderson, Donald A. 30
Bennett, Michael P. 85
Boardman, Thomas J. 15
Brown, David L. 49
Bugbee, A.D. 73
Burdick, Richard K. 51
Christensen, Howard B. 36
Cooke, William P. 30
Faulkner, J. Earl 89
Folks, J. Leroy 41
Gillings, Dennis B. 46
Graham, Jack E. 76
Hahn, Gerald J. 15,80
Hill, William J. 15
Hocking, Ronald R. 15
Hunter, William G. 15
Isenberg, Daniel J. 67
Lackritz, James R. 59
Lawton, William H. 15
Little, Charles H. 85
McGovern, Paul G. 91
May, Lawrence E. 76
Merrill, H.D. 73
Minton, Paul D. 6,28
Oberg, Robert J. 54
Ott, R. Lyman 15
Rencher, Alvin C. 1
Rubison, Richard M. 57
Selbert, W.F. 8
Sheikh, Habeeb A. 76
Sherr, Lawrence A. 89
Shiffler, Ronald E. 59
Snee, Ronald D. 15
Starry, A.R. 8
Strahan, Robert F. 91
Strawderman, William E. 15
Ware, James H. 44
Warren, J.A. 73
Watts, Donald G. 61
Zahn, Douglas A. 67

Note: The papers and discussions in the Proceedings volumes are reproduced exactly as received from the authors. None of the papers has been submitted to a refereeing process. These presentations are presumed to be essentially as given at the Annual Meeting. The 1979 Proceedings volumes are not copyrighted by the Association; hence, permission for reproduction must be obtained from the author, who holds the rights under the copyright law.

Authors in these Proceedings may submit their papers to any journal of their choice. The ASA Board of Directors has ruled that publication in the Proceedings does not preclude publication elsewhere.

TABLE OF CONTENTS

I. STUDENT EVALUATION OF INSTRUCTION

Chairman: H. GILL HILTON, Brigham Young University

An Analysis of Student Evaluation of Instruction and Relationships Among Salary, Chairman's Ratings, and Student Ratings of Faculty - ALVIN C. RENCHER, Brigham Young University 1

Administrative Use of Student Surveys - One View - PAUL D. MINTON, Virginia Commonwealth University 6

Students' Evaluations: Their Antecedents and Aftermath - W.F. SEIBERT and A.R. STARRY, Purdue University 8

II. PREPARING STATISTICIANS FOR CAREERS IN INDUSTRY

Chairman: DONALD G. WATTS, Queen's University

Report of the Committee on Training of Statisticians for Industry - RONALD D. SNEE, E.I. du Pont de Nemours & Co., Chairman; THOMAS J. BOARDMAN, Colorado State University; GERALD J. HAHN, General Electric Company; WILLIAM J. HILL, Allied Chemical Corporation; RONALD R. HOCKING, Mississippi State University; WILLIAM G. HUNTER, University of Wisconsin; WILLIAM H. LAWTON, Eastman Kodak Company; R. LYMAN OTT, Merrell Research Center; WILLIAM E. STRAWDERMAN, Rutgers University 15

Discussion - PAUL D. MINTON, Virginia Commonwealth University 28

III. TEACHING SERVICE COURSES IN STATISTICS

Chairman: LARRY J. RINGER, Texas A&M University

A Modular Self-Paced Statistics Course: Five Years' Experience - DONALD A. ANDERSON and WILLIAM P. COOKE, University of Wyoming 30

Student's Choice -- Self-Pacing and/or the Lecture Method for Introductory Statistics - HOWARD B. CHRISTENSEN, Brigham Young University 36

The Elementary Statistics Course -- Fact or Fiction - J. LEROY FOLKS, Oklahoma State University 41

IV. PRINCIPLES AND PROBLEMS IN TEACHING STATISTICS - TARGETED TO MEDICAL SPECIALISTS

Chairman: THEODORE COLTON, Dartmouth College

Teaching Statistics to Medical Specialists - Cardiologists - JAMES H. WARE, National Heart, Lung, and Blood Institute 44

Discussion - DENNIS B. GILLINGS, University of North Carolina at Chapel Hill 46

GENERAL CONTRIBUTED PAPERS

A Course in Elementary Inferential Statistics That Students Understand - DAVID LILE BROWN, University of Connecticut 49

On the Use of Canonical Analysis to Test MANOVA Hypotheses - RICHARD K. BURDICK, Arizona State University 51

An Empirical Approach to Teaching General Statistics Augmented by the Computer - ROBERT J. OBERG, Framingham State College 54

A Probability Model for Multiple Choice Tests - RICHARD M. RUBISON, Kansas State University 57

A Teacher's Aid for the Binomial Probability Distribution: A College Football Contest - RONALD E. SHIFFLER, Georgia State University and JAMES R. LACKRITZ, San Diego State University 59

A Task-Analysis Approach to Designing a Regression Analysis Course - DONALD G. WATTS, Queen's University 61

Non-Statistical Aspects of Statistical Consulting - DOUGLAS A. ZAHN and DANIEL J. ISENBERG, Harvard University 67

CONTRIBUTED PAPERS - POSTER SESSION

Statistical Test Item Collection System - A.D. BUGBEE, H.D. MERRILL and J.A. WARREN, University of New Hampshire 73

On Selecting Best Predictor Questions in a Standardized Pass-Fail Examination - JACK E. GRAHAM, LAWRENCE E. MAY and HABEEB A. SHEIKH, Carleton University 76

A Guide to Texts on the Design of Experiments - GERALD J. HAHN, General Electric Company 80

Cases: A System of Case Studies for Teaching Statistics - CHARLES H. LITTLE, St. John's University and MICHAEL P. BENNETT, Xaverian High School 85

Transparency Masters for Characteristics of Probability Distribution and the Normal Probability Distribution - LAWRENCE A. SHERR, University of Kansas and J. EARL FAULKNER, Brigham Young University 89

What's in a Number? Verbal Equivalents of Numerical p-Values - ROBERT F. STRAHAN and PAUL G. McGOVERN, Iowa State University 91

AN ANALYSIS OF STUDENT EVALUATION OF INSTRUCTION AND RELATIONSHIPS AMONG SALARY, CHAIRMAN'S RATINGS, AND STUDENT RATINGS OF FACULTY

Alvin C. Rencher, Brigham Young University

Introduction

A sizable literature has appeared on the subject of student ratings of teacher effectiveness. Even a partial list of references would easily exceed the length of the text of this paper and will, therefore, be omitted entirely. We mention only a few general claims and counter-claims:

1) Students (especially underclassmen) are too inexperienced to make consistent judgements; similarly, students cannot make accurate judgements until several years after they took the course. Others claim there is no significant difference in ratings of underclassmen and upperclassmen and that alumni give ratings similar to those of currently enrolled students.
2) Only chairmen or experienced colleagues with publication records are in a position to make a valid evaluation of instruction vs. the claim that student ratings are positively correlated with chairman and colleague ratings.
3) Student ratings are not related to the rank or degree of the instructor vs. the claim that instructors holding higher degrees tend to be rated higher.

Widespread disagreement among faculty members about the merits of student ratings would seem to point up the fact that the concept of teaching effectiveness is a complex phenomenon to evaluate. There will undoubtedly be a continuing effort to monitor and improve student ratings.

Objectives

An organized student rating scheme has been in use at Brigham Young University for about twenty years. This report presents the results of two statistical studies made at different time points.

The original survey consisted of four sections: (1) 10 questions about the student, (2) 22 questions about the course and instructor, (3) a checklist of 30 ways the instructor could improve the course, and (4) additional space the students could use to write their own comments. Most items were similar to those used elsewhere in student evaluation instruments.

An abbreviated version of the survey follows:

Section 1 Information about the Student

1. Class standing (freshman, sophomore, etc.)
2. Age
3. Sex
4. Church affiliation
5. Number of courses (previously taken in the department)
6. Grade average
7. Class grade (expected)
8. Subject matter (like or dislike)
9. Student preparation (hours per week)
10. Purpose (major, elective, etc.)
11. (Previously deleted)

Section 2 Information about the course and instructor

12. Objectives
13. Organization
14. Work load
15. Suitability of text
16. Scholarship
17. Attitude
18. Preparation
19. Interest of class
20. Personal help
21. Stimulation
22. Tolerance
23. Participation
24. Clarity
25. Achievement
26. Evaluation
27. Grading
28. Spiritually inspiring
29. Laboratory
30. Effectiveness
31. Recommendation
32. Rating of survey itself
33. Personal reaction to filling out survey

Section 3 Checklist -- 30 ready-made suggestions the student can check to help the teacher improve the course.

In 1972, the University Faculty Committee on Teaching contemplated a revision and sought statistical answers to the following questions: (1) which of the 22 questions in Section 2 are most redundant and could be deleted with least loss of information, (2) is the checklist (Section 3) useful, and (3) which questions best separate faculty members in a ranking scheme?

Data

The data available for this study consisted of several past administerings of the survey, both university wide and selected. For the major purpose, individual student responses were preferable, and it was found that only the survey results from Fall 1971 furnished this kind of data. The other data sets consisted of class summaries where individual responses had been amalgamated into averages for the entire class. Inasmuch as interest centers on an individual student's (multivariate) response to the various questions, it was necessary to use

the Fall 1971 data sets to obtain the information needed. The other data sets were used, however, in obtaining information on other aspects of the investigation.

The number of surveys completed in the Fall of 1971 was 38,000. A 25% sample of these was selected since processing of the entire data set would have involved a prohibitive amount of computer time.

Correlations

The first information obtained was in the form of correlations between responses to questions. Although interest was centered on Section 2, correlations of responses from Section 1 among themselves and with the questions in Section 2 were also obtained. The highest of these was a correlation of .70 between questions 1 and 2 which is hardly surprising. Also noteworthy were correlations of .52 between questions 1 and 5 and .41 between 2 and 5. No further analysis was performed for the first ten questions.

The individual correlations between questions in Section 2 showed only a few of substantial size. The largest correlation was .79 between questions 30 and 31. This correlation is large enough to justify deleting one of these two questions. The second highest correlation is .63 between questions 19 and 31 followed by correlations of .62 between 19 and 30 and .62 between 32 and 33. These are not large enough to immediately suggest deletion of any of these questions without further investigation of their relationships with other questions.

Certain other individual correlations are of interest. For example, a significant negative correlation between the responses to question 33 and either 30 or 31 might be expected. This would indicate that a student who liked a teacher would resent the time spent on the survey, while the student who disliked a teacher would welcome the opportunity to fill it out. This expectation did not materialize, however, as a small positive correlation appeared instead of a negative one. The correlations of .07 between 30 and 33 and .06 between 31 and 33 show that students' reaction to the survey was essentially independent of their rating of the instructor.

In order to determine which questions in Section 2 were the best candidates for deletion, further analysis beyond simple correlation coefficients was pursued in order to consider each question as it relates to all the others collectively in addition to how it relates to each of the others singly. A question may not be extremely highly correlated with any other single question and yet its response might be satisfactorily predicted from knowledge of all the other questions. The information it contains may already be available in several other responses considered together.

Deleting Variables

Among the 22 questions in Section 2, there are 231 different pairs. An attempt to evaluate and compare this many correlations may not reveal the true patterns of relationship. Various methods of deleting variables have been suggested (see Joliffe, 1972, 1973) involving regression, principal components, and cluster analysis. The technique used in this case involved factor analysis.

Consider the usual (orthogonal) factor analysis model for p variables expressed in terms of m factors.

$$\begin{aligned} x_1 &= \lambda_{11}f_1 + \lambda_{12}f_2 + \ldots + \lambda_{1m}f_m + e_1 \\ x_2 &= \lambda_{21}f_1 + \lambda_{22}f_2 + \ldots + \lambda_{2m}f_m + e_2 \\ &\vdots \\ x_p &= \lambda_{p1}f_1 + \lambda_{p2}f_2 + \ldots + \lambda_{pm}f_m + e_p \end{aligned} \qquad (1)$$

In matrix form (1) can be expressed as $\underset{\sim}{x} = \underset{\sim}{\Lambda}\underset{\sim}{f} + \underset{\sim}{e}$. The following will be assumed: $E(\underset{\sim}{x}) = \underset{\sim}{0}$, $\mathrm{cov}(\underset{\sim}{x}) = \underset{\sim}{\Sigma}$, $E(\underset{\sim}{f}) = \underset{\sim}{0}$, $\mathrm{cov}(\underset{\sim}{f}) = \underset{\sim}{I}$, $E(\underset{\sim}{e}) = \underset{\sim}{0}$, $\mathrm{cov}(\underset{\sim}{e}) = \underset{\sim}{\Psi} = \mathrm{diag}(\Psi_1, \ldots, \Psi_p)$, and $\underset{\sim}{e}$ and $\underset{\sim}{f}$ are independent. From these assumptions it follows that

$$\underset{\sim}{\Sigma} = \underset{\sim}{\Lambda}\underset{\sim}{\Lambda}' + \underset{\sim}{\Psi} \; . \qquad (2)$$

In practice the correlation matrix is often used in place of $\underset{\sim}{\Sigma}$.

If two variables x_i and x_j have nearly identical loadings on the m factors, they contribute similar information and one of the two may be deleted. The following weighted distance function is suggested as a measure of the degree of closeness or similarity of x_i and x_j:

$$d_{ij} = w_1(\lambda_{i1} - \lambda_{j1})^2 + w_2(\lambda_{i2} - \lambda_{j2})^2 + \ldots + w_m(\lambda_{im} - \lambda_{jm})^2 \; . \qquad (3)$$

If the λ_{ik} are estimated by factoring $\underset{\sim}{\Sigma}$ (or $\underset{\sim}{\Sigma}$-$\underset{\sim}{\Psi}$) using eigenvalues and eigenvectors, appropriate weights are provided by the eigenvalues. Thus, w_1 is the largest eigenvalue, w_2 is the next largest, etc. If another method has been used to estimate the loadings or if they have been rotated, appropriate weights can be found from

$$w_k = \sum_{i=1}^{p} \lambda_{ik}^2 \qquad k = 1, 2, \ldots, m$$

The following procedure is suggested to determine which variables can best be discarded. Find d_{ij} for each pair x_i and x_j, $i \neq j$. Arrange these in ascending order. Then from the pair of x's with the smallest distance function value, (3), one can be deleted. Similarly, one of the two variables with next smallest d_{ij} can be deleted. This process can continue until as many variables as desired are discarded. If two variables are found to be similar and later each is found to be close to a third variable, two of the three may be deleted. This suggests the possible use of a clustering procedure. However, cluster analysis is not recommended due to the problem to be discussed next which can best be handled in the pairwise framework.

Negative correlations have an effect on the location of points in the space of factor loadings. If the response on a question is given on a multipoint ordered scale between poles, a reversal of the poles will change the sign of the correlation of the given question with all others. From (2) it is clear that

$$\sigma_{ij} = \sum_k \lambda_{ik}\lambda_{jk}, \quad i \neq j .$$

A change in sign produces

$$-\sigma_{ij} = \sum_k (-\lambda_{ik})\lambda_{jk} = \sum_k \lambda_{ik}(-\lambda_{jk}) .$$

Thus, either $(\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{im})$ or $(\gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jm})$ has all signs reversed and appears on the opposite side of the origin from its original position. If two points are close together in the factor loading space, a reversal in sign of the correlation between the two variables they represent would place the two points very far apart. The distance function d_{ij} would then fail to detect the close similarity between the two variables. Allowance can be made for this possibility when making each comparison by considering $\sum w_k(\lambda_{ik} + \lambda_{jk})^2$ as well as $\sum w_k(\gamma_{ik} - \gamma_{jk})^2$ in each case and choosing the smaller as the distance between the two variables in question. Thus (3) becomes

$$d_{ij} = \min \left\{ \sum_k w_k(\lambda_{ik} - \lambda_{jk})^2, \; \sum_k w_k(\lambda_{ik} + \lambda_{jk})^2 \right\} . \qquad (4)$$

If the number of variables to be discarded has not been predetermined, some visual assistance in arriving at an appropriate value can be obtained by ordering the distances and plotting them. The plot can then be examined for a turning point where a noticeably steeper ascent begins.

A factor analysis was performed on the correlation matrix for the 22 questions in Section 2. It was found that 17 factors were required to explain 95% of the variation. The distance function (4) was calculated for each of the $\binom{22}{2} = 231$ pairs of variables. The smallest 20 of these are given in Table 1. The remaining distances ranged up to a high of 2.875.

TABLE 1

Pairs of questions with smallest distance functions

Question Pair	Distance	Question Pair	Distance
30,31	.023	25,30	.657
32,33	.041	13,30	.797
19,31	.151	13,31	.877
19,30	.170	16,30	1.016
12,27	.172	19,25	1.072
13,18	.174	18,30	1.082
19,21	.352	16,31	1.112
21,30	.453	13,25	1.113
21,31	.473	13,19	1.132
25,31	.603	24,30	1.163

Note that the pair of variables that are closest, 30 and 31, are also both very close to 19. Thus, only one of these three variables need be retained.

As an aid to determining how many questions might appropriately be deleted, the first 15 distances are plotted in Figure 1. There is a definite change in pattern from the sixth to seventh value. It seems clear that a variable can be deleted from each of the first two pairs with almost no loss of information. The next four distances also appear small enough to warrant deletion of one variable from each pair if desired. Beyond that point, however, further deletion may not be justified.

Ranking the Faculty

Previously, question 30 had been considered the most valid over-all measure of teaching effectiveness. Other questions which have been used to rank the faculty or identify the top 10% are numbers 16, 17, 18, 19, 21, and 31. In an attempt to determine statistically which questions actually serve best to rate and distinguish faculty, a principal components analysis was performed. Using the coefficients in the first principal component, it was found that the following questions serve best to separate teachers. These are in descending order:

30, 31, 19, 21, 25, 13, 12, 18, 24, 20, 26, 16, 17, 28, 27, 22, and 23.

Check List

An attempt was made to statistically evaluate the utility of the check list (Section 3 in the survey) in past surveys. If a

professor uses the check list to improve his teaching, it would be expected that the following survey would show an improvement in his rating and a corresponding decrease in the number of responses to the above list. A program was prepared to search past experiences to find out if this occurrence had been observed with any degree of regularity. The following five surveys were examined: Spring 1967, Fall 1968, Fall 1969, Spring 1970, and Fall 1970. The computer searched for those instructors who appeared more than once in those five surveys. It was discovered that 1134 faculty members and graduate teaching assistants had been surveyed more than once. The average number of checks per student was recorded along with the average response to question 30. These two numbers were then compared for each survey in which an instructor appeared. A decrease in the number of checks accompanied by an increase in rating would appear as a negative correlation, as would also the opposite, i.e., an increase in the number of checks with a corresponding decrease in rating on question 30. A correlation was calculated for each faculty member and the average was found for all those in the list. This average correlation was found to be -.37. This figure, though small, is statistically significant and indicates that the average number of checks per student and the average rating do indeed tend to move in opposite directions.

Correlation of Student Rating with Rank and Salary

This part of the study was conducted more recently. Based on the earlier study, the survey was revised and the revised version was used in the later study.

The most recent campus-wide administering of the survey was in 1976. Data were obtained on rank, salary and experience for the same time period. There were 734 faculty members for whom such data were available. There were four values of rank corresponding to Instructor, Assistant Professor, Associate Professor, and Professor. Salary was for the academic year and experience was simply the total number of years spent in college teaching. A single student evaluation rating was obtained for each teacher by averaging over all appropriate items on the questionnaire and over all students taking the survey for that teacher. The following correlations were obtained:

	Correlation between student rating and	
1.	Rank	.091
2.	Experience	-.150
3.	Salary	-.094

With $n = 734$ these are all statistically significant ($p \leq .014$). However, they are too small to be of any practical importance and the results essentially support the null hypothesis of independence between student rating and each of the three variables.

Correlation Between Student Rating and Chairman Rating

For 178 of the faculty members in the study, an evaluation by the department chairman was available. This evaluation consisted of a rating on each of five scales. Each scale was accompanied by another scale showing the value to the department of the corresponding type of activity. These value levels were used as coefficients in computing a weighted average of the five rating scales. The correlation between the department chairmen rating and student rating was .254 ($p = .0006$, $n = 178$). This correlation is larger than those for rank, experience, and salary but it is still rather small ($r^2 = .0645$). There appears to be little relationship between a faculty member's value to the department as perceived by the chairman and the teacher effectiveness as perceived by the student.

In an attempt to relate student rating to all four variables, rank, experience, salary, and chairman rating, the data were submitted to a stepwise regression program. Only the two variables, chairman rating and rank, entered into the model with resulting R^2 value of .087 ($p = .0003$, $n = 178$). The gain in R^2 from .0645 to .087 is statistically significant ($F = 4.33$, $p = .039$) but .087 is still not of sufficient magnitude to claim that a relationship of any practical significance has been demonstrated.

As a further effort to improve the fit, various transformations (powers, products, logs) were performed on the four dependent variables. In a stepwise regression on these transformed dependent variables, R^2 achieved a value of only .095 which is not a significant increase from .087.

REFERENCES

JOLLIFFE, I. T. (1972). Discarding variables in a principal component analysis. I: Artificial data. Applied Statistics, 21, 160-173.

JOLLIFFE, I. T. (1973). Discarding variables in a principal component analysis. II: Real data. Applied Statistics, 22, 21-23.

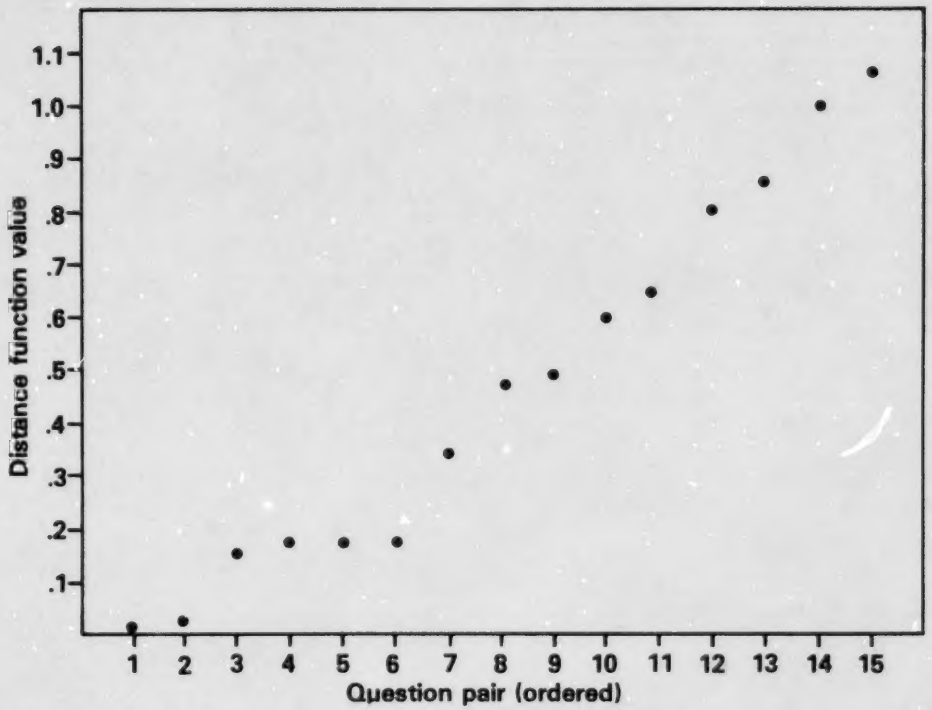


Figure 1. The Fifteen Pairs of Questions with Smallest Distance Function Values

ADMINISTRATIVE USE OF STUDENT SURVEYS - ONE VIEW

Paul D. Minton, Virginia Commonwealth University

Let me begin by mentioning some resources for those interested in this subject. There is a rapidly growing literature on faculty evaluation and now a quarterly newspaper, Faculty Development and Evaluation in Higher Education. The AAUP Policy Documents and Reports includes a "Statement on Teaching Evaluation," and the AAUP publishes a booklet, The Recognition and Evaluation of Teaching, which offers an extensive bibliography.

Involving students in the evaluation process is fairly new, although experiments with student feedback have been carried on for many years. Attempts to achieve consensus on procedures for official overall evaluation of faculty are not very old, and my impression is that we are just at the stage of modest beginnings of development of such procedures, in which student ratings of some kind will undoubtedly play a part.

An article in the summer, 1978 issue of the newspaper entitled "Faculty Evaluation," by Gaff, Festa, and Gaff may be a good summary of where we are. It mentions several student rating forms which have come into widespread use and which you may want to know about. There is a discussion of questions that have been raised about student ratings and a summary of the answers the authors have found in the literature. The questions concern whether students are reliable reporters, and what factors affect ratings (e.g., popularity, amount of learning). Literature summaries, the authors say, indicate general agreement among students as to effectiveness of instructors, persistence of student perceptions after graduation, and relative independence of ratings from circumstances thought to inspire bias, such as sex, course grade, or year in school. As to the amount of learning being reflected in evaluations, the evidence is contradictory - "One would hope to rate teachers by the amount of learning taking place, but this ideal seems unlikely of achievement in the near future." The studies also show that in spite of shortcomings, student evaluations still serve important purposes in giving students a chance to affect behavior of professors, allowing faculty to analyze aspects of their teaching, and providing one component of of an overall evaluation procedure. The authors note that "there is no evidence that faculty or administrative judgments of teaching quality are more reliable than students' judgments." They note also that systematic evaluation procedures may be the best guarantee for faculty that their interests are protected. Student opinion is only one of several sources that should be mentioned in comprehensive evaluation - the teacher in question, colleagues and administrators are also sources, but with distinctive roles: faculty can assess the instructor's knowledge via course objectives, the syllabus, etc., and students can best judge enthusiasm, skill, technique, etc. Teaching outside the classroom, e.g. advising, needs assessing. Regularity of assessment can identify progress, stability or regression of performance. It is helpful for faculty to be able to compare results with other instructors. It is in the interest of the entire professoriat and of the institution to make evaluations used for personnel decisions on the best available evidence. Whether evaluation improves teaching is still unknown.

The AAUP "Statement" is in somewhat more general terms, but is in general agreement. Student learning is the most valid indicator of teaching performance, but is most difficult to measure. Research supports the practical usefulness of data from trained observers, faculty colleagues and students. Student perceptions are a prime source of information from those who must be affected if learning is to take place. Faculty members should be meaningfully involved in any systematic efforts to obtain student opinion. It is important to obtain data over a range of teaching assignments and over a period of time. Evaluations that go only to the professor may be of use in improving his performance, but contribute little to the process of faculty review. Classroom visitation over a period of time may be useful, if carefully done; having colleagues visit does not yield reliable data. The limitations of self-evaluation are obvious; this cannot be the sole method of evaluationn, but can still be helpful. Faculty members should have a primary though not exclusive role in evaluating an individual faculty member's performance as a teacher.

Evaluation of College Teaching, by Grace French - Lazovik, published as "An Occasional Paper" of the Association of American Colleges in 1975, is another good summary of the literature. The author feels that "excellent advice is being ignored - for those who would heed it, start with Kenneth Eble's The Recognition and Evaluation of Teaching, the AAUP Statement on the Evaluation of Teaching, and R. I. Miller, Developing Programs for Faculty Evaluation." She feels that the literature has made it clear that it will never be possible to isolate the teacher's contribution from all of the other variables acting - they are numerous, complex, subtle and non-observable. Thus, other criteria must form the basis of judgment. There are four main sources - self-evaluation, peer evaluation based on classroom visitation, peer evaluation without visitation and student evaluation. She stresses that evaluation should never rely on a single source. She summarizes findings from the literature as to the strengths and weaknesses of each source, as to purposes (summative and formative evaluation), and as to requirements for effective implementation. Students can play two roles - reporter and evaluator: they can report many things but can only evaluate certain things, and a good evaluation process will allocate such roles and interpretations appropriately.

My experience has been consistent with the general picture described above of the status of

faculty evaluation and the role of student surveys in it. That is, that faculty evaluation is difficult, that there is not much agreement as yet in academe on how it should be done, and that student evaluations are helpful, but only a part of the story. Even how much a part they should play is still being debated. Processes for full evaluation of teaching, research and other forms of service to the institution are being developed and improved; they will have subjective elements, will involve student, colleague and administrators' feedback, and will take into account the mission, goals and objectives of the institution and of the department.

The use of student evaluations can be improved as we work together constructively on the problem. There is noise in them, and there are still disagreements as to the meaning of experiments that have been performed to answer various questions of interest to us, for example, whether student ratings are correlated with grades. For some time, it seemed that they were not. Then a study seemed to indicate that they are negatively correlated, and the author concluded that students rated most highly those from whom they learned the least. The experimental design of that study has been severely criticized. Another study concluded that student ratings are positively correlated with grades, and interpreted that as proof that student ratings measure effectiveness of learning. Another paper argued that the reason we use student evaluations is to relieve administrators of the responsibility of exercising judgment about teaching performance. (This is the paper that states that there is no correlation between quality of teaching and quality of research, and that the degree of correlation varies from institution to institution.)

Thus, it is clear that much research needs to be done in this area, and with better experimental design and analysis. It is not clear who should or will do this research, but the level of interest and activity is rising. A fairly recent phenomenon is the campus office of faculty development, designed to assist faculty members in improving their teaching effectiveness. Ours at V.C.U. has been very successful; perhaps since evaluation is one aspect of improvement, these offices may take the lead in research on evaluation.

Improvement of the evaluation process and of decision-making based on evaluation will come as faculty and administration work together constructively on the problem. If the use of student ratings has produced grade inflation, as has been charged, the conclusion to draw may not be to discontinue student ratings but to improve the reward system and decision-making structure, by according the student ratings a more appropriate role or by a change in procedure. We have asked faculty to annotate student ratings with any pertinent facts, for example.

In summary, we are today at the stage of evaluating faculty by compiling information from as many sources as feasible, including student ratings, and over as long a period as possible, and then trying to judge the value of each piece of information, and looking for corroboration and weight of evidence. There is noise and subjectivity in all of it. We are trying to improve the whole system by eliminating noise where possible and by understanding and fairer interpretation where noise cannot be eliminated. The way is open for us as faculty to work on how we want to be evaluated, but not, of course, to reject evaluation.

Students' Evaluations: Their Antecedents and Aftermath

W. F. Seibert and A. R. Starry

Purdue University

Last winter, when I agreed to take part in this session and then needed a title for these remarks, I finally evolved the one printed in the program, "Students' Evaluations: Their Antecedents and Aftermath." It seemed like a good idea at the time. But I suspect now that that title is more provocative and, to use the jargon of psychology, more of a projective technique than it is informative, so perhaps I should take a minute or two to justify it and provide some related background.

From the beginning, the chairman's invitation had included the suggestion that I give particular attention to the effects of ratings - general effects on students' behavior, more specific effects on students' learning, and effects on teacher behavior or teacher performance. The immediate difficulty with that suggestion was that the literature of the field is largely silent on these matters. It offers a couple of plausible but not well confirmed ideas concerning general effects on students, it says nothing about ratings' effects on students' later learning (or, for that matter, it says nothing about the effects of ocean tides on lunar motion), and it has barely begun to look seriously at the effects which ratings may have on later teacher performance.

There are, for example, some who contend that students will be motivated to evaluate more carefully if they evaluate relatively early in the semester or term, i.e., at a time when instructional shortcomings might be corrected soon enough to benefit the students directly. Similarly, there is a widely held suspicion that, if students were asked to evaluate every course each semester, they would soon become "saturated" and their judgments would lose credibility and sensitivity. In neither case is there convincing evidence supporting these conclusions, but both still have some plausibility. Also, it is a long-standing and an often unspoken assumption that rating results constitute an effective first step toward the improvement of teaching and there are ample anecdotes in support of this view. Recently, several studies have tried to put this idea to the test, but with mixed results.

In outline, these studies typically provide for two or more administrations of an instructional rating instrument, perhaps at mid-semester and at semester's end. Half of the participating teachers are given immediate access to their mid-semester results and half are not. Also, in some studies, the first group of teachers may receive some form of help in interpreting the ratings and in developing modifications of their teaching style. In any event, most of the half dozen published studies show little or no change occurring in the treated or experimental group; however, since there is an exception or two to this typical result and since there are several substantial ways to improve upon the study procedures typically used, it seems too early to pronounce any firm verdict. For now, probably the most that can be said is that the value of ratings in improving teaching (or, more accurately, in improving later ratings) is not as great or not as easily demonstrated as investigators had expected.

In any case, because there are gaps in the literature and lore of ratings which I needed somehow to navigate around, my counter-proposal to our chairman's suggestions was that I would spend the allotted time examining several of the common themes and issues which, in our experience, precede institutional adoption of rating systems, as well as those which may follow adoption. In creating this before-and-after dichotomy, I do not mean to say that the division is as sharp as I have implied or that each issue must inevitably arise. Rather, I would like to think that I am describing a reasonably common and, for the most part, a healthy or a hopeful scenario.

Before proceeding further, it seems only fair that I also indicate something about my background and biases. Briefly, the background is this: Seven years ago, several colleagues and I devised the rationale, then the procedures and materials to implement a new and fairly novel instructional rating system. As we had hoped, the system was quickly adopted by increasing numbers of faculty, first locally, then on a number of other campuses, and, currently, it appears to be used more or less routinely at 75-100 colleges and universities. For present purposes, what seems most important is that we have now corresponded or worked directly with those institutions and perhaps with two hundred others as each one undertook to reach local decisions concerning the uses, policies, or desired characteristics of an instructional rating system. These institutions include almost every variety: they are large and small, public and private, some are research-oriented and some teaching institutions, and they include citadels of liberal education as well as those with more vocational or professional curricula. Also, in our contacts with them, their questions have covered the full range, from the most specific and practical to the most sweeping and philosophical. In short, then, as we have interacted with involved faculty and administrators, we have had the chance to observe how institutions

grapple with questions which student ratings pose and, frequently, to see what their decisions are and how they are implemented.

To summarize my own biases concerning ratings, I should probably confess to the following: First, they are, I believe, an acceptable way to call increased attention to teaching and, for many institutions, they may also function to shift the balance toward undergraduates and toward the classroom, while at the same time moderating the influence of grants and contracts and publications. Secondly, ratings seem the most practical means of obtaining reliable assessments of course and instructor quality and, technically, they are at least as satisfactory as alternative methods, such as colleague ratings, administrator ratings, self-ratings, classroom visitation, and direct assessments of student learning. Third, ratings hold some promise - but just how much, we can't yet say - as a means of monitoring specific features of instruction, then devising improvements, wherever these seem needed. Fourth, when the results of ratings are used thoughtfully, they deserve a place alongside other information which enters into promotion, salary, or tenure decisions. If this sounds radical, it is not meant to. Instead, it is based largely on the conviction that such decisions should incorporate information from every relevant and accessible source. Finally, and perhaps idiosyncratically, our experience during these last seven years persuades me that a flexible evaluation system, one which essentially allows and encourages each teacher to create his or her own rating form, is superior to a system which provides only one form (or a few) for use in all circumstances. Traditionally, most rating instruments reflect an implicit assumption that undergraduate teaching is quite homogeneous, at least in its methods and instructional goals. Even if the assumption were correct, which I doubt, faculty should be free to decide which aspects of a course are important and, thus, what the contents of an evaluation should include. For those faculty who teach traditionally, the customary rating items and rating forms may be satisfactory, but for those who depart from the norm, perhaps by adopting a self-paced format, by team teaching, computer-based simulations, or other non-traditional methods and goals, evaluations should be adapted to reflect the circumstances, as they exist.

As we consider current methods and uses of student ratings, it is easy to imagine that these are recent additions to the methods and means used in higher education and in one sense they are, since ratings have gained much of their prominence in the last decade or two; however, their history reaches into the 1920's and thus their use has been examined and debated for at least a half century. Nevertheless, questions and issues which emerged early have in many cases stayed late and they can still arouse as much controversy as ever. Also, as implied earlier, the issues and questions run the gamut from fundamental matters of educational philosophy or purpose to technical and procedural details. For example, at one extreme, there are some who argue against ratings on the grounds that they require students to perform in an improper role and, in the process, they undermine sensitive and "correct" relationships between teachers and students. At an opposite extreme, some faculty are simply interested to know how many student raters are needed to achieve acceptable stability or reliability of results. But even though the range of issues and questions is broad, it seems to me that most of the recurring questions deserve the attention they receive. More often than not, they reflect a healthy concern and they function as an antidote for oversimplifications or for complacency about instructional evaluation or about our grasp of the dimensions or determinants of instructional effectiveness. At the very least, we can benefit from recurring reminders that instruction is a process too complex and too individual ever to be captured or represented by a few numbers or a few words. The evaluation results available to us may be informative, they may be credible, and they may even have demonstrable validity, but their use still requires the most careful consideration and interpretation.

My purpose during these past few minutes has been to establish some groundwork for a further consideration of student ratings, and especially for a consideration of certain stages and issues or questions which often precede institutional adoption of ratings, as well as later issues which can follow adoption and the implementation of these services. Since several of the things I'm about to say may suggest that most colleges and universities have highly similar dynamics, evaluation needs, and decision-making processes, I should first deny that that has been our experience. Instead, although some institutions seek packaged and pat answers to their (or their faculty's) evaluation needs, it rarely seems advisable to encourage these efforts. For me, the two most apparent reasons for proceeding more deliberately are, first, that institutional climates can differ markedly and, thus, a "transplanted" evaluation system with borrowed and transplanted policies may be rejected by the new host or recipient institution. The situation resembles that of organ transplants between two distinctive individuals. The second reason, if I may stretch that analogy a bit, is that institutions, like individuals, seem to have "immune systems" which function to reject foreign bodies. Thus, with institutions as with individuals, the immune system needs to be understood and treated and, in some crude sense, suppressed if a successful transplant is to be achieved. We have witnessed and have been impressed, for example, by the distinctive climates, needs, and prospects at three institutions with which we have had a fair amount of contact. The first is a medium-sized community college with a unionized faculty and a fresh history of intense faculty-administration discord. The next is a small Quaker college which, not surprisingly, has virtually the opposite tradition. And the third is a particularly large public university which has a

strong, even a distinguished faculty, one which is also noted for guarding its turf jealously and effectively. In any case, among institutions as disparate as these, no one can expect that a single or simple set of transplanted policies could possibly be responsive to each institution's needs.

In discussing how institutions decide on instructional ratings and their use, I expect to sound at least a little envious and, in fact, it is difficult not to envy those institutions, several of which we have observed and worked with, which muster their best planning skills and devote them to the creation and selection of local procedures and policies. Such action is, I think, especially enviable because it demonstrates the institution's seriousness about the importance of teaching and demonstrates its desire to foster and maintain improvements. At the same time, such institutions show a concern for the long-term future and not simply for the "quick fix."

Among those institutions we know best and, more specifically, among those which seem to evolve successful solutions, there is a slightly idealized composite institution and institutional process that I'd like to describe. The process we've observed goes something like this: First, within the institution a shared awareness evolves that current evaluation practices are unsatisfactory and need to be changed. Following preliminary discussions, a small committee is then formed and assigned the task of developing a recommended course of action. My guess is that the best committees probably include not only a highly credible and respected faculty member who is known also as a skillful teacher, but they include a spokesman or advocate of faculty rights, an administrator, perhaps a student, and some person with the formal background to examine the related literature critically. We also find that computer center representatives are often included in committees that we see, but perhaps because our own system is computer-based and its implementation requires the involvement of computer facilities and operations staff.

In any event, once a committee is formed, it will normally require at least six months and, more likely, a year to investigate local needs and interests and to study practices and trends at other institutions, then to prepare its recommendations. Parenthetically, it seems to me that the single act of establishing a committee signifies that two important bridges have been crossed. First of all, it provides some assurance that new evaluation practices will not suddenly appear following a decree from on high. Whatever plans are eventually adopted, they will at least be influenced and perhaps will directly result from a thorough analysis of alternatives and from group efforts. Secondly, by establishing a committee, the institution is, as I see it, acknowledging that the decisions to be made are not simple and that they reach well beyond the gateway question: "Shall we have a new evaluation system? Yes or No?"

The more specific and more frequent questions that we have found certainly include questions of cost, but the dominant interests are more basic and serious than that. In any event, costs are likely to average twenty to thirty cents per student evaluator or, roughly five to six dollars per classroom, although they can be considerably higher. Among the more serious questions, perhaps the foremost is that of mandatory versus voluntary faculty participation. Not infrequently, institutions are under some pressure from boards or legislatures or lay citizens, so they need a carefully considered response to the question. In our own case and in many others that we know, participation is essentially voluntary, even though subtle pressures are sometimes present and operating to restrict true freedom of choice. Also, if participation is to be mandatory, how often will faculty participate? Once a year? Once a term? Or perhaps in every course they teach? Surely, the answer to the last question is "no," so the principal question becomes one of sampling adequately among courses and across time.

The question of confidentiality also arises frequently. Will evaluation results be provided to the faculty member above or can others have direct access to them? We have elected to operate as a confidential service which deals directly with participating faculty and we have never regretted that decision. In fact, I strongly suspect that much of the early success on our campus is attributable to our policy of confidentiality. Still, I can imagine circumstances where strict confidentiality is not a critical factor; it is pleasant to imagine, too, since it probably reflects that, at the institution, there is a high level of mutual trust and shared purpose.

There is one other prominent question which often precedes the adoption decision, then there are some secondary matters I'd like to mention briefly. The prominent - and fairly complicated - question concerns the principal purpose which ratings are to serve. The two main options are first, to encourage faculty to seek student feedback and, in this way, to monitor and perhaps improve their teaching, while the other is to use ratings as one element in deciding promotion and salary questions; and the caveats here are many. As we will indicate later, it is extremely doubtful that one rating system can serve both purposes satisfactorily. Also, if the chosen purpose is to encourage instructional improvement, what supporting or technical or consulting services will faculty receive? Or, if the second purpose is chosen, what steps will be taken to assure the integrity of the results and to assure that ratings are obtained under standard and dependable conditions?

And finally, what accommodations will be made for faculty who teach unconventional or non-traditional courses? Must they seek evaluations of the quality of their lectures or of the course text, even though neither of

these is used? Perhaps they could be better judged on the basis of their tutorial efforts or their use of simulations. And what about norms? Should ratings be interpreted directly or by reference to norms and percentile equivalents? And if norms are used, should they be undifferentiated or, instead, developed on the basis of similar courses and of faculty in similar circumstances? Last of all, the very practical questions involved with inaugurating a new service sometimes arise. The best general answer is surely to start small and to allow at least one semester, and preferably two or more, for faculty and operations staff to gain familiarity and confidence without unnecessary pressure.

Our own Purdue experience and much that we have observed on other campuses suggests that several issues are likely to emerge and, perhaps, to recur once a rating system has been instituted. For me, the issue I was least prepared to expect is that of pressure for student access to evaluation results, usually in order to publish them and, as students like to say, to "encourage better teaching." I can see now that it was naive not to expect this development, but I suspect that several things deceived us. First, as you may recall, student activism in the 1960's often included prominent complaints about undergraduate teaching. Activists deplored the wide use of graduate assistants; they deplored professors whose interests were said to be centered in research, grants, publications and graduate teaching; and they intended to do whatever they could to make faculty more responsible for providing effective undergraduate teaching.

A common course of action on many campuses was to establish a student-run instructional rating system and to publish the results once or twice a year. But with the arrival of the 1970's, activism declined, student interest in preparing for a career increased, and besides, students seemed to grow tired of the work involved in operating an evaluation and reporting system. In addition, many institutions had taken some corrective steps to meet students' earlier complaints, including instructional development and faculty development programs and the establishment of institutionally-sanctioned rating or evaluation systems. So, many of the student-run systems declined and died. It appeared they had had their day, served their purpose, and would now fade away. However, I now suspect that student interest in gaining access to faculty and course ratings will continue to ebb and flow for the foreseeable future. Where students think they can obtain existing data, they may try that, but if not that, then part of the history of the 1960's could repeat itself.

We have noticed, too, that, once in place, ratings often come to be regarded as if they were extremely precise, rather than only numerical representations of some complicated and often subjective judgments. I doubt that this is a unique problem affecting only instructional ratings, but it is nevertheless troublesome. For reasons I can't fathom, colleagues who are otherwise sensible and sensitive will treat rating results almost as if they were carefully calibrated laboratory measurements. For them, a rating of 4.2 is surely better than a 4.1 or a 50th percentile is only average while a 52nd percentile is above average. And in extreme cases, some have been known to sum several ratings for each of several persons, then use the totals to rank individuals and even to grant or withhold rewards. In any event, I am convinced that continuing and forceful efforts are needed to educate those who would treat ratings simplistically. We need, I believe, to stress the general message that effective teaching, like scholarly accomplishment, is extremely difficult to define and to gauge.

A further issue and problem can probably be referred to as "erosion of original intent" and there seem to be at least two important ways in which erosion can take place. First, for those systems which are intended to help faculty monitor and improve their teaching, these can gradually be transformed into systems which serve largely to enhance promotion and salary decisions. To this non-economist, the process resembles Gresham's law. On the one hand, we have laudable and idealistic efforts to create authentic improvements in instruction, but on the other, individual faculty have the possibility of prompt recognition and reward. If, by some means, the faculty member can construct a credible case that he or she is a good or excellent teacher, then immediate rewards and benefits are likely to follow. And there is a great likelihood that that strategy will seem more attractive to faculty under pressure than will the strategy of laboring over course improvements, with no assurance that these efforts will finally be successful. It has been carved in stone that the price of liberty is eternal vigilance and much the same thing should be said about preserving and rewarding sincere efforts toward improved instruction.

The second form that erosion may take involves the transformation of voluntary systems into mandatory or almost mandatory systems. The pressures and the dangers seem less than in the case just mentioned, but they still are bothersome. What may happen is this: A system with voluntary participation is inaugurated and its results eventually become generally accepted as credible; besides, there is little or no competing information to be found, so a partial vacuum of information exists. One thing leads to another until faculty who used to think it was voluntary are now encouraged, then strongly encouraged to participate. These are not easy pressures to resist.

I have said nothing about the validity issue, but it clearly is one which cannot be ignored. It must be the most persistent and among the most difficult to address. There are perhaps a few hundred validity coefficients reported in the literature and they stem from studies that have at least their fair share of

procedural weaknesses. The customary study plan involves several sections of a lower division course, several instructors, a common final examination administered to all sections, and a common instructional rating form. What, then, is the Pearson r between average ratings, section by section, and average examination scores? I know of no general or summarizing answer to that question because the results of existing studies have varied widely, but the two most careful studies which I know, one by John Centra and the other by Sullivan and Skanes, include several estimates, derived from a variety of courses and disciplines, and the median r from both studies is about .40. Perhaps the more interesting feature of these results is that they have little consistency, as you compare across courses or disciplines. Admittedly, the N's are generally small and so some inconsistency is to be expected, but I would doubt that that accounts for the entire problem. I cannot pretend to make sense out of the median values or the inconsistencies, but I have begun to wonder whether students' judgments are not generally more valid indicators of effective teaching than the modal study suggests. This may be wishful thinking, but my suspicions stem in part from the observation that studies often show some restriction of range on the variables under study, they rarely take account of unreliability in their measures, they cannot deal effectively with instructors' varied interpretations of ostensibly common or shared instructional goals, and I know of no attempts to deal with the widely held belief that individual teachers often attain their effectiveness - or their failures - along idiosyncratic paths.

Finally, I should mention that it can prove unusually difficult to modify and improve an evaluation system, once it has been planned and implemented. Thus, there is one more reason to plan carefully at the outset. Once a service is established, faculty and others develop certain routines and expectations relating to the service and they may, as in our own case, have certain supplies or materials that serve to define the procedures of participation. These things can create a considerable inertia and they may delay or even prevent the introduction of otherwise desirable changes.

(For other but similar purposes, we recently prepared a fairly extensive bibliography of studies and other publications relating to student ratings of instruction. It is appended here for those who wish to examine that literature.)

References

[1] Aleamoni, L. The usefulness of student evaluations in improving college teaching (Research Report #356). Urbana: Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1974.

[2] Aleamoni, L. Typical faculty concerns about student evaluation of instruction. Paper prepared for presentation at the Symposium on Methods of Improving University Teaching, The Technion-Israel Institute of Technology, Haifa, March 1974.

[3] Astin, A. & Lee, C.B.T. Current practices in the evaluation and training of college teachers. In C.B.T. Lee (Ed.), Improving college teaching. Washington, D.C.: American Council on Education, 1967.

[4] Bassin, W. M. A note on the biases in students' evaluations of instructors. Journal of Experimental Education, 1974, 43, 16-17.

[5] Bejar, I. I. A survey of selected administrative practices supporting student evaluation of instructional programs. Research in Higher Education, 1975, 3, 77-86.

[6] Borgatta, E. Sidesteps toward a nonspecial theory. Psychological Review, 1954, 61, 343-352.

[7] Borgatta, E. Student ratings of faculty. AAUP Bulletin, 1970, 56, 6-7.

[8] Braunstein, D., Klein, G., & Pachla, M. Feedback expectancy and shifts in student ratings of college faculty. Journal of Applied Psychology, 1973, 58, 254-258.

[9] Centra, J. A. Evaluating teachers. Letter in Change, 1973, 5, 7.

[10] Centra, J. A. Effectiveness of student feedback in modifying college instruction. Journal of Educational Psychology, 1973, 65, 395-401.

[11] Centra, J. A. Student ratings of instruction and their relationship to student learning. American Educational Research Journal, 1977, 14, 17-24.

[12] Costin, F., Greenough, W., & Menges, R. Student ratings of college teaching: Reliability, validity, and usefulness. Review of Educational Research, 1971, 41, 511-535.

[13] Doyle, K. O., Jr. Student evaluation of instruction. Lexington, MA: D. C. Heath & Co., 1975.

[14] Drucker, A. J. & Remmers, H. H. Do alumni and students differ in their attitudes toward instructors? Journal of Educational Psychology, 1951, 42, 129-143.

[15] Eble, K. Professors as teachers. San Francisco: Jossey-Bass, 1972.

[16] Eble, K. Letter in The American Scholar, 1972, 41, 494-495.

[17] Frey, P. Student ratings of teaching: Validity of several rating factors. Science, 1973, 182, 83-85.

[18] Frey, P. The ongoing debate: Student evaluations of teaching. Change, 1974, 5, 47-48; 64.

[19] Gage, N. L. The appraisal of college teaching. Journal of Higher Education, 1961, 32, 17-22.

[20] Gessner, P. Evaluation of instruction. Science, 1973, 180, 560-570.

[21] Guthrie, E. R. Measuring student opinion of teachers. School and Society, 1927, 25, 175-176.

[22] Guthrie, E. R. The evaluation of teaching: A progress report. Seattle: University of Washington, 1954.

[23] Heilman, J. D. & Armentrout, W. D. The ratings of college teachers on ten traits by their students. Journal of Educational Psychology, 1936, 27, 197-216.

[24] Hildebrand, M., Wilson, R. C., & Dienst, F. R. Evaluating university teaching. Berkeley: University of California Press, 1971.

[25] Hoyt, D. & Howard, G. The evaluation of faculty development programs. Kansas State University Research Report #39. Office of Education Research, 1977.

[26] Hudleson, E. The validity of student ratings of instruction. School and Society, 1951, 73, 265-266.

[27] Kerlinger, F. Student evaluations of univerity professors. School and Society, 1971, 99, 353-355.

[28] Kossoff, E. Evaluating college professors by "scientific" methods. The American Scholar, 1971-72, pp. 79-93.

[29] Kulik, J. A. & McKeachie, W. J. The evaluation of teachers in higher education. In F. N. Kerlinger (Ed.), Review of research in education (Vol. 3). Itasca, IL: Peacock, 1975.

[30] Lovell, G. D. & Haner, C. F. Forced-choice applied to college faculty rating. Educational and Psychological Measurement, 1955, 15, 291-304.

[31] McKeachie, W. J. Student ratings of faculty. AAUP Bulletin, 1969, 55, 439-444.

[32] McKeachie, W. J. Use of student ratings in evaluation of college teaching. National Institute of Education, Grant No. NE-G-00-3-0110 Final Report, March, 1975.

[33] McKeachie, W. J., Lin, Y-G., & Mann, W. Student ratings of teacher effectiveness: Validity studies. American Educational Research Journal, 1971, 8, 435-445.

[34] Maslow, A. H. & Zimmerman, W. College teaching ability, scholarly activity and personality. Journal of Educational Psychology, 1956, 47, 185-189.

[35] Menges, R. J. Letter in Change, 1973, 5, 7.

[36] National Project III. Evaluation of college teachers (Criteria Report No. VI). Ann Arbor: Center for Research on Learning and Teaching, University of Michigan, 1977.

[37] Pambookian, H. S. Initial level of student evaluation of instruction as a source of instructor change after feedback. Journal of Educational Psychology, 1974, 66, 52-56.

[38] Powell, R. Faculty rating scale validity: The selling of a myth. College English, 1978, 39, 616-629.

[39] Remmers, H. H. The relationship between students' marks and student attitude toward instructors. School & Society, 1928, 28, 759-760.

[40] Remmers, H. H., Shock, N. W., & Kelly, E. L. An empirical study of the Spearman-Brown formula as applied to the Purdue rating scale. Journal of Educational Psychology, 1927, 18, 187-195.

[41] Remmers, H. H. Rating methods in research on teaching. Handbook of research on teaching (Chapter 7). American Educational Research Association. Chicago: Rand, McNally & Co., 1963.

[42] Rodin, M. Can students evaluate good teaching? Change, 1973, 5, 66-67; 80.

[43] Rodin, M. & Rodin, B. Student evaluations of teachers: Students rate most highly instructors from whom they learn the least. Science, 1972, 177, 1164-1166.

[44] Seibert, W. F. National Project III: Elevating the importance of teaching (Fund Associate's Final Report to Fund for the Improvement of Postsecondary Education, Grant No. G007502063). West Lafayette, IN: Measurement and Research Center, Purdue University, 1977.

[45] Seldin, P. How colleges evaluate professors. New York: Blythe-Pennington, Ltd., 1975.

[46] Spencer, R. E. & Aleamoni, L. M. A student course evaluation questionnaire. Journal of Educational Measurement, 1970, 7, 209-210.

[47] Sullivan, A. & Skanes, G. Validity of student evaluation of teaching and the characteristics of successful instructors. Journal of Educational Psychology, 1974, 66, 584-590.

[48] Vattano, F. J. Letter in The American Scholar, 1972, 41, 495-496.

PREPARING STATISTICIANS FOR CAREERS IN INDUSTRY

Report of the Committee on

TRAINING OF STATISTICIANS FOR INDUSTRY

American Statistical Association
Section on Statistical Education

COMMITTEE MEMBERS

Ronald D. Snee, E. I. du Pont de Nemours & Co., Chairman
Thomas J. Boardman, Colorado State University
Gerald J. Hahn, General Electric Company
William J. Hill, Allied Chemical Corporation
Ronald R. Hocking, Mississippi State University
William G. Hunter, University of Wisconsin
William H. Lawton, Eastman Kodak Company
R. Lyman Ott, Merrell Research Center
William E. Strawderman, Rutgers University

ABSTRACT

This report provides guidelines for universities to consider in developing programs for training statisticians who will work in industry. Useful information for students who are considering industrial employment is also included. The recommended programs focus on real problems and the statistical theory and methodology which are useful in their solution. Technical competence is only one of many factors which industry considers when hiring and promoting statisticians. When a statistician leaves school his or her skills and experiences should include statistical knowledge, practical problem solving, consulting practice, and the ability to communicate orally and in writing with nonstatisticians. There are many details which must be worked out (eg, content of specific courses and organization of consulting internship programs) and it is hoped that the statistical societies and universities will form committees, hold conferences, and develop programs to further address these issues. Many of our recommendations apply more broadly to the training of all types of practicing statisticians.

FOREWORD

At the 1978 American Statistical Association (ASA) meetings in San Diego the ASA Section on Statistical Education formed a Committee on "Training of Industrial Statisticians" comprised of both industrial and academic statisticians. The committee's objective was

> "To make recommendations on training of statisticians for industrial employment at the BS, MS, and PhD levels, with emphasis on conveying the perceived needs of employers to academic institutions."

This report summarizes the committee's work on this important area of statistical education. The major emphasis is on programs at the MS and PhD levels. We discuss not only appropriate training programs but also related issues such as characteristics of successful industrial statisticians and desirable faculty backgrounds and interests. This material will, therefore, be useful not only to university faculties, but also to students of statistics who desire information on the necessary background for, and nature of, careers in industry. This report focuses on the training of statisticians for industry (also referred to as industrial statisticians), specifically those working in the physical, biological, chemical, and engineering sciences. We believe, however, that most of the views expressed are applicable to the training of all types of practicing statisticians. We feel that similar studies should be made for the training of statisticians in other areas (eg, government and business statistics).

1. INTRODUCTION AND SUMMARY

The purpose of this report is to discuss appropriate training for statisticians who wish to pursue careers in industry and to raise the level of awareness of opportunities in this field. We recognize that industry provides one of several career opportunities for statisticians and that few universities can afford a program designed specifically for this application area. We further recognize that few departments can provide all the programs that we identify and that the faculty and graduates of these programs cannot have all the characteristics which we list as desirable. We also point out that much of what we recommend is being done in varying degrees at several universities. Our intent is for the contents of this report to help these departments create even better programs and to provide a reference point for other universities which wish to initiate this type of training.

We believe that the opportunities for statisticians in industry have not been fully recognized. The growth of this field depends on the number of well-trained statisticians selecting this career. A characteristic of all successful industrial statistics groups is statisticians who take leadership roles in the solution of company

problems and the development of statistical methodology. These individuals bring problems to the attention of management, develop solutions, and, through personal persuasiveness, see that their recommendations are implemented. They understand the role of statistics in solving industrial problems and are able to communicate this role to scientists, engineers, and managers as well as to their colleagues.

This type of statistician will most likely be developed by a university department which provides the student with a balanced view of theory and practice. The theory taught must include that which is needed to understand and extend or tailor the statistical methodology useful in practice. The courses on statistical methods should concentrate on techniques which are useful in solving problems subject to the constraints of the real world. One should be careful to avoid teaching "cookbook applications" which result in an unrelated set of tools with no overall integrating theme. By the same token, the theory courses should contain the theoretical philosophy which underlies the statistical methodology and avoid the teaching of "cookbook theory" which reduces to a disconnected collection of theorems and lemmas.

The importance of having a balance between theory and practice can be seen by drawing an analogy between a university program and the functions of a corporation. The vertices of the figures in Figure 1 represent the functions of the two organizations. Each vertex (function) is connected (interacts) with all other vertices. We believe that the interaction of teaching, research, and consulting (ie, statistical practice) affects the quality of a university statistics program in the same way as the cooperative efforts of the research, development, manufacturing, and marketing organizations affect the performance of a corporation. We note further that the marketing organization depends on research, development, and manufacturing for the products and services it sells and provides input to these organizations concerning the needs of the consumer. In a similar fashion, universities' theoretical and applied research and teaching functions produce the practitioners of the profession and the theory and methodology they use in their work. The practitioners in turn can provide input on the training and statistical methodology which is needed. It is important in both organizations to catalyze these linkages in such a way that their output is best suited to the marketplace.

This Committee has studied many aspects of industrial statistics (Figure 2) and has developed guidelines for university statistics departments to consider in the formation of programs to train statisticians for industry. Our recommendations, summarized in Table 1, should be viewed in the light that programs developed along the indicated lines will best prepare statisticians for careers in industry. It will be apparent that a few new courses will be needed. It may also be appropriate to reexamine the content and objectives of some of the existing courses. Our recommended programs focus on real problems and the statistical theory and methodology which are useful in their solution.

The issues surrounding our recommendations are discussed in the following sections. We begin by summarizing what industry looks for in a statistician. We then discuss what universities can do to help students prepare for careers in industry. Next we describe areas for cooperation among university, industry, and the statistical societies. The report concludes with a detailed summary of our conclusions and recommendations.

TABLE 1

SUMMARY OF CONCLUSIONS AND RECOMMENDATIONS FOR THE TRAINING OF STATISTICIANS FOR INDUSTRY

- Provide a balance of theory and practice.
- Include a wide variety of statistical techniques.
- Discuss the realities of statistical practice.
- Give students hands-on experience with real problems.
  - Consulting intern programs are very effective.
- Improve communication skills.
  - Report writing and oral presentations can be included.
  - Teaching experience is valuable.
- Illustrate statistical methodology with real problems.
- Provide computer experience.
- Focus on training at both the MS and PhD levels.
- Faculty interest in problem solving is a key factor.
- The faculty's attitude toward industry is important.
- A "Careers in Industrial Statistics" brochure is needed.
- Universities and industry can assist each other by working together to create appropriate training programs.
- Programs to further address these issues can be developed jointly by the universities and the statistical societies.

2. WHAT DOES INDUSTRY LOOK FOR IN A STATISTICIAN?

2.1 The Industrial Statistician - A Profile

In developing new programs and modifying existing programs for the education of industrial statisticians it is important to understand the characteristics, responsibilities, and backgrounds of

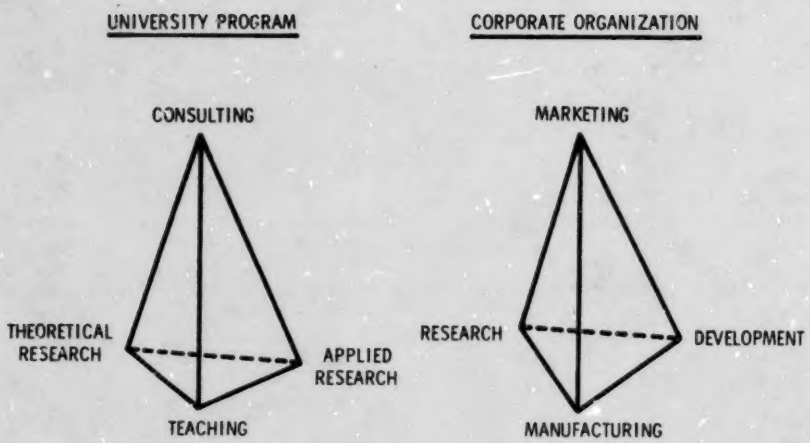


FIGURE 1

COMPARISON OF THE ORGANIZATIONAL FUNCTIONS OF UNIVERSITIES AND CORPORATIONS

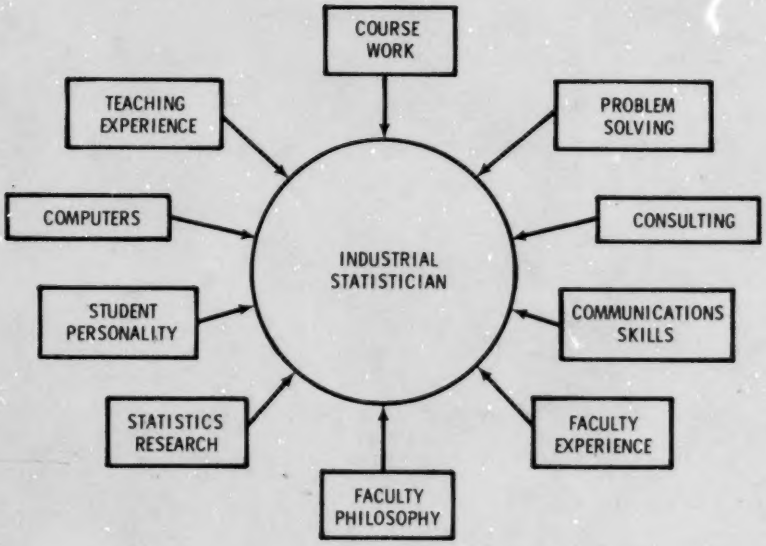


FIGURE 2

TRAINING OF INDUSTRIAL STATISTICIANS – THE CRITICAL COMPONENTS

those who work in this field. The characteristics of an effective industrial statistician are summarized in Table 2. Personal characteristics are discussed in this section. Those related to university training will be discussed in the following sections of this report. This information may be useful to faculty in identifying those students who have these characteristics and informing them of the opportunities in this field. This material will certainly be helpful to students who wish to consider a career in industry.

Industrial statisticians have interests that combine business, science, problem solving, statistics, and computing. They come from a wide variety of backgrounds including all fields of science, engineering, and mathematics. The common characteristic seems to be a strong interest in working on practical problems. Many are tired of studying but get excited about using statistics to solve real problems and get great satisfaction from helping others (Daniel, 1969).

Industry looks for those individuals who can begin to function immediately and have potential for continued professional growth. We believe that such growth will be enhanced if their training and outlook is oriented toward the use of statistics in problem solving and contains a balance of theory and applications. Industrial statisticians gain recognition through the quantity, quality, timeliness, and impact of their work; "do it, do it well, do it now, see that it is used". Good oral and written communication skills are essential to accomplish these ends. Much of the industrial statistician's time is spent in defining the problem and in educating others about what can and cannot be accomplished with statistics.

Personality is another key consideration. Effective industrial statisticians are usually confident, outgoing, optimistic, enthusiastic, tactful, and organized. Industry also looks for statisticians who are creative, inquiring, resourceful, skeptical, good listeners, and willing to work under pressure. They must be prepared to work on several problems at the same time and be able to adapt quickly to new and different problems or challenges. Motivation, high energy level, and ability to do independent work are also key characteristics.

Broadness of view and ability to see the big picture are necessary because of the diversity of problems and situations that one will later encounter. It is important to have the ability to see the big picture and to focus on the statistical methodology that is needed to solve the problem at hand. The industrial statistician often has to iterate between these two extremes many times during the solution of a problem.

TABLE 2

CHARACTERISTICS OF AN EFFECTIVE INDUSTRIAL STATISTICIAN

- Is well trained in the theory and practice of statistics.
- Is an effective problem solver.
- Has good oral and written communication skills.
- Can work within the constraints of the real world.
- Knows how to use computers.
- Is familiar with the statistical literature.
- Understands the realities of statistical practice.
- Has pleasing personality and is able to work with others.
- Gets highly involved in the solution of company problems.
- Is able to extend and develop statistical methodology.
- Can adapt quickly to a new problems and challenges.
- Produces high quality work in a timely fashion.

3. WHAT TRAINING WOULD BE MOST HELPFUL FOR STUDENTS PREPARING FOR CAREERS IN INDUSTRY?

3.1 The Course Work Should Take Into Account That the Industrial Statistician Is a General Practitioner and Should Contain a Balance Between Theory and Practice

Most statisticians in industry are necessarily generalists because they are presented with problems from a wide variety of subject matter areas and must employ many different statistical tools in solving them. It is not unusual for a statistician working in a research and development laboratory of a chemical company, for example, to be asked to evaluate animal toxicity studies, industrial hygiene data, or studies in clinical chemistry. Similarly a biostatistician working for a

pharmaceutical company may be involved with clinical studies of several drugs as well as preclinical studies related to biochemistry, pharmacology, and toxicology. It is impossible for all statisticians to be exposed to all useful statistical techniques during their training. What is needed, however, is the recognition of this aspect of statistical practice and for the educational process to be designed accordingly. Briefly stated, the student should be exposed to

- Statistical theory and methodology useful in the solution of a wide range of real problems, and
- A philosophy to use as a guide in applying these methodologies and developing others as needed in solving problems.

This concept not only gets the students off to a good start but also provides them with a proper framework for future growth and development.

The specific content of statistics courses is too large a subject to give full justice to in this report. We decided, therefore to identify the following statistical techniques which are used most often by industrial statisticians:

Statistical Technique	Use By Industry
Regression analysis	Routinely by most companies
Basic statistical methods	
Analysis of variance	
Graphical display & data summary	
Design of experiments	
Variance component estimation	Usually depends on area of application
Biostatistics including bioassay	
Categorical data analysis	
Quality control & acceptance sampling	
Nonlinear estimation	
Nonparametric methods	
Ranking/paired comparison data analysis	
Multivariate analysis	
Probability modeling	Less frequently
Simulation	
Reliability and life data analysis	
Numerical analysis	
Time series analysis	
Survey sampling	

These subjects are listed in "approximate" order of frequency of use overall and within each of the three use-categories. This order is based on our experiences and an informal survey we made of 20 industrial statistics groups. The point is that the list is long and that it is important to have some appreciation of the utility and limitations of each of these methods. One possibility might be to have an "appreciation course" or a series of "miniseminars" which would make students aware of the statistical literature in fields which they do not study in detail.

Students preparing for careers in industry at the MS level should have courses in design of experiments, regression, analysis of variance, analysis and interpretation of data, and sampling. Other courses should be selected depending on the student's intended area of employment. The courses which form the core of the MS program are also important at the PhD level and will be studied in greater detail. At this advanced level of training there is also the opportunity to select from a wider variety of course offerings.

We believe that more statistics programs should provide instruction in the areas of quality control, acceptance sampling, response surface methodology, and reliability and life data analysis. Quality control and acceptance sampling techniques are used in all industries. Response surface methodology is particularly useful in the process industries (eg, chemical and food). The electronics industry makes considerable use of reliability and life data analysis techniques.

The curriculum built around any subset of the field of statistics should contain a balance between theory and practice. The level of theory is dictated by what is needed to understand the statistical methodology which is useful in practice and to extend existing methodology when its assumptions are overly restrictive. The mathematical level is dependent on the theory level. Let there be no mistake about it, relevant and useful theory are an essential part of the education of an industrial statistician.

Courses on consulting techniques and effective communication are also important. In some instances these needs can be provided by a consulting intern program if a faculty member with consulting experience (preferably in industry) provides the instruction (see Section 3.2.1).

It is also essential that students be knowledgeable in at least one programming language (eg, Fortran, Basic, APL) and be familiar with most of the popular statistical packages (BMDP, MINITAB, OMNITAB, RUMMAGE, SAS, SPSS, etc). It is likely that early in his or her career. Depending on the circumstances, statisticians may be asked to do their own programming, use canned computer programs, work with a professional programming staff, or some combination of these. In any event a knowledge of statistical computing will make this part of the job go much smoother.

Students should receive instruction in proper numerical analysis techniques for statistical computation in addition to computer programming training. Many existing computer programs use poor numerical analysis programming techniques and can produce erroneous results. Some institutions may choose to combine the programming and numerical analysis into a single course.

3.1.1 It is Important to Include the Realities of Statistical Practice

We feel that the training of industrial statisticians does not require the addition of many new courses. It may be appropriate, however, to reexamine the content and objectives of some of the existing courses. The "realities of statistical practice" can be included in the course work as well as other aspects of a statistics program (computer programming, consulting courses, internship programs, etc). The following illus-

trate the type of considerations we have in mind.

Problem Formulation & Analysis. It is important to teach the process by which a problem is formulated, the experimental design constructed (if appropriate), the data collected, the analysis specified and completed, and the conclusions developed. The most important, and often the most time consuming, part of a statistical analysis is the proper definition of the problem. The various aspects of the problem solving process may be difficult to include in all courses; however, the student should be informed as to when and where in the program these issues will be addressed.

Graphical Displays. The single most effective statistical tool is a graphical display. Exploratory and summary plots should be mandatory in all courses involving data. (We know of at least one instructor who emphasized this point by requiring all "surprise quizzes" be taken on graph paper with a penalty assessed if none was available.) Plotting often provides an adequate solution to a problem and frequently provides at least 50-75% of the information which is needed.

Historical Data rarely contain as much information as it would appear to those not familiar with this type of data. Teachers should make sure that students understand both the strengths and the weaknesses of analyses of historical data (see Box, Hunter, and Hunter, 1978, pp. 487-498).

Statistical Literature. The industrial statistician relies heavily on the statistical literature. This should be reinforced in the course work, pointing out to students that keeping up with recent developments in the field and knowing how to find things in the literature (journals, textbooks, authors, leaders in the field) are responsibilities of a professional statistician.

Adequate Analysis. At appropriate points in different courses instructors should emphasize the need for determining how much analysis is needed. We noted earlier that often a plot of the data provides an adequate solution. Students should be reminded that scientists are more impressed by the usefulness of what you do for them than by how sophisticated your statistical analysis is.

This is only a partial list. Each faculty will want to create its own list in consultation with other academic and industrial statisticians, scientists, and engineers. Many of these realities will be appropriate for inclusion directly in courses. Others will be appropriate for inclusion in our recommended general course on "Analysis and Interpretation of Data" (see Section 3.2.2). Some will be covered in consulting courses and intern programs. Students will learn more about the realities of statistical practice the "hard way" as they practice their profession.

3.2 Students Need Hands-On Experience with Real Problems

We believe that hands-on experience with real problems is an essential ingredient and should be part of the program of all statistics students. This experience should be obtained both from training in statistical consulting and from courses on the design of experiments and analysis of data. A PhD thesis which solves a real problem and involves considerable data analysis can also be an effective mechanism (see Section 3.5).

3.2.1 Consulting Intern Programs Are Very Effective

Training in statistical consulting can be obtained via

- internship programs,
- working in consulting laboratories,
- taking consulting courses, and
- interacting with other disciplines through joint seminars and rap sessions.

We believe that working as a consulting intern, with close supervision by an experienced statistical consultant, is the best way to obtain hands-on experience because of the realistic environment. One gains experience not only in problem formulation, data analysis and interpretation, but also communication of the results, and project planning and organization. All of these skills are essential to success as an industrial statistician. There are two other important benefits. It will be apparent to both the faculty and the student whether the student has, as yet, the necessary interpersonal communication skills (eg, personality, patience, tact, cooperative attitude) to be a successful consultant and the student will get some direct experience on what life as a practicing statistician is like.

Working with a nonstatistician also helps develop communication skills. The translation of the statistical results to the language and needs of the client is part of the exercise. Statistical results are of little value if the client doesn't understand them and put them to work. The success of an industrial statistician is a direct function of the impact of his or her work on company business.

It is appropriate for students to participate as interns at both the MS and PhD levels. While the program will vary with the student, it is beneficial for students to work 1-2 semesters as a consulting intern at the MS level and 1-2 years at the PhD level. In these programs the student begins by participating in meetings with the client, doing the necessary computing work, and summarizing the results for presentation to the client. As the student gains experience this role expands to the point that the student begins to function independently.

The internship program may present problems for some universities because of the lack of experienced faculty and, in some instances, the statistics department may not be set up to provide consulting services. It will be difficult to involve part-time students in consulting intern programs at the university; however, some of these students will have less need for this type

of training because they will get some of these experiences in their regular occupation. There may also be the tendency (which should be avoided) to lengthen the program to include this type of training. Efforts to overcome these problems are justified, however, because the students who have this experience are more likely to be better practicing statisticians.

Some statisticians, particularly those from industry, have expressed concern that university consulting may not properly model industrial consulting. It is pointed out that a significant fraction of the work is "one or two shot consultation" on MS and PhD theses and the client is never seen again. This is a valid concern to which universities should give some attention; however, university and industrial consulting have some things in common. The industrial statistician is frequently called on to do spot consulting, often over the telephone. This activity is similar to consulting on MS and PhD theses. There is another more important parallel; that of developing a consulting business. In those situations where the statistics faculty and consulting interns are doing a good job, the subject matter thesis advisor generally gets involved in the discussions and a working team develops: statistician, consulting intern, thesis advisor, graduate student. It is not uncommon for one or more consulting interns to work on 2 or 3 theses associated with the same research project over a several year period. An active consulting business and a long association involving many good problems develop. Most of the consulting skills developed are directly applicable in the industrial environment. The principal aspects of industrial consulting which are missing in the university are the focus on the business impact of the statistical contributions and the frequent need to produce results under tight deadlines.

Experience with real problems can also be obtained through a cooperative program in which the student spends time in industry working with an experienced statistician. This arrangement is attractive because of the industrial environment. It may not be possible to implement this approach in some areas, however, because of lack of industrial statistics groups. It may also be difficult to use this approach to train a large number of students. This type of internship would certainly be worth exploring in those instances where the necessary interest and expertise exist.

In some consulting labs students are trained via intern programs under the supervision of experienced statisticians; however, in others statistics graduate students provide assistance to all who seek help with little supervision from the faculty. This arrangement presents problems for both the statistics student and the client. The students are left to their own devices to develop the necessary consulting skills and often do not develop a correct understanding of the project they are working on. The client may get poor advice and, worse yet, gets a false impression of what contributions statistics can make. A golden opportunity to develop a "believer" is lost, sometimes forever.

Consulting courses are useful, particularly when taken in conjunction with an intern program. These courses are limited, however, because they attempt to develop consulting skills in an unrealistic setting. There is little client contact and few opportunities to work on a project from beginning to end which results in limited hands-on experience. Statisticians are ready to be directly exposed to problems midway through their masters program and should be given this opportunity.

Joint seminars with other disciplines are useful to identify new problem areas and initiate dialogues. This method is generally ineffective in maintaining long-term associations unless the identified problems are worked on jointly and solutions developed with the scientists. Some scientists may be less likely to describe their problems to a group of statisticians than they are to a single statistician. Some people are reluctant to seek help from others fearing that to do so is a sign of weakness.

There is a growing literature on the subject of consulting and problem solving. The reader is referred to Marquardt (1979) and Woodward and Schucany (1977) and their extensive bibliography for additional information on this topic.

3.2.2 Real Problems and Associated Data Can Benefit Many Courses

Hands-on experience with real problems can also be obtained by including real data sets in courses. The use of each statistical technique should be illustrated with a real data set and sufficient problem background should be given so that it will be clear how the statistical analysis contributed to the solution of the problem. This can be done in many courses including those on probability.

Students should also be required to collect and analyze their own data as part of at least one course. This is best accomplished by having each student plan, conduct, and analyze an experiment as discussed by Hunter (1977). Another approach, which has been used effectively for a number of years by Andrews (1979), is for the student to search the subject matter literature (eg, SCIENCE, TEXTILE RESEARCH JOURNAL, NATURE, TAPPI, etc.) for an experiment with real data which is reported with little or no statistical analysis. The student is required to comment on the experimental design and reported analysis, perform the appropriate analysis, and write up the results. Both of these approaches have the useful by-product of producing a never-ending supply of real problems and data sets which can be used in the teaching of other courses. The latter approach has the limitation that the student is not able to discuss the problem with the scientist who collected the data.

In addition, there should be a one or two semester course on "Analysis and Interpretation of Data" which is based totally on real problems. This course would begin to develop many of the skills (problem formulation, data analysis and "snooping", presentation of results, report writing) which are needed to work on real problems. Sufficient problem background would be

provided so that the student would see both the problem and the analysis, and how statistical methods are often modified in practice to develop an adequate (not necessarily optimum) solution. The book, PROCESS QUALITY CONTROL (Ott, 1975), was developed from the notes for such a course which E. R. Ott taught at Rutgers University for many years.

We also feel that it is very important that statisticians who are going into health related industries have instruction in the monitoring and the execution of studies (ie, monitoring of the quality of study data). Students should also be involved with the construction and maintenance of large data base files and the documentation of data trail from receipt of original data through various stages of the analysis. Knowledge of data base management systems is becoming increasingly important in a number of areas (eg, environment and energy). These experiences can also be obtained in a consulting internship program.

3.3 Improvement of Communication Skills Is a Valuable Addition to a Program

Good oral and written communication skills are essential to the industrial statistician and are important factors in both the hiring and advancement of industrial statisticians. Promotion to higher technical levels and supervisory positions requires especially good oral and writing skills. The inability to communicate effectively, particularly with nonstatisticians, is a major deficiency of a large number of statistics graduates. This may also be true in other fields; however, it is critical for statisticians because of the multidiscipline nature of the work. Communication skills can be developed through one-on-one consulting (Section 3.2.1), course projects, seminars, special courses, and teaching.

The development of communication skills can be part of the entire graduate program. One mechanism is the assignment of course projects and MS and PhD papers which involve the writing up and presentation of a project related to a real problem. The report should contain a clear summary of the significance of the work which is free of statistical jargon. The technical details should appear in an appendix. It would also be helpful for the students to give 10-minute presentations for a nonstatistical audience. Communication skills can also be improved by the preparation and presentation of reports and critiques of published research (Section 3.2.2) in written form and seminars. This approach is most effective when individual guidance is provided by the student's mentor.

Universities should consider introducing a special course on communication skills perhaps in cooperation with the Speech and/or English departments. It has been suggested that this course be taken in lieu of the foreign language requirement. Courses or seminars that include "role playing" are helpful in teaching communication skills. This technique could be used by having a student present an assigned topic to one or more other students who play the role of the nonstatistical client. A critique is provided at the end of the presentation. Emphasis on communication skills should be part of both regular courses and special courses or sessions. Instructors will have a positive impact if they set a good example by making their classes informative and interesting.

3.3.1 Teaching Experience is Important

Industrial statisticians spend a significant portion of their time teaching small groups as part of their consulting, making presentations to client groups, and teaching in-house statistics courses. The ability to teach, therefore, is a valuable asset. In addition to improving communication techniques, teaching

- Improves one's understanding of the subject,
- Builds self-confidence in appearing before a group,
- Enables one to learn to handle difficult and penetrating questions,
- Helps one to organize thoughts and explain difficult concepts, and
- Provides an opportunity to organize and manage a project.

It is beneficial for students to have 1-2 years of teaching experience preferably teaching basic statistics to nonstatisticians.

3.4 Programs Should be Geared to the MS Level and Higher

Most companies have several (5-7) different technical ranks and pay levels for statisticians. Some typical job titles and entry degree levels are shown below.

Position Title	Entry Degree
Assistant Statistician	BS
Associate Statistician	MS
Statistician	PhD
Senior Statistician	
Statistics Consultant	
Senior Statistics Consultant	

Some may prefer the title "statistical scientist" over statistician since the emphasis is on scientist with the word statistical indicating the type of scientist.

The different job titles provide paths of progress and reflect different levels of competence. While there are jobs at the BS level (eg, assistant statistician, computer programmer), it is generally agreed that the MS degree is a minimum requirement for the professional statistician. There are always exceptions but it is recommended that someone interested in statistics as a profession obtain solid foundations in science or engineering and mathematics. A science or engineering major should be sure to take a large number of mathematics courses. Those majoring in mathematics should have a strong minor in science or engineering. Knowledge of basic science is valuable in communicating with scientists and can

be acquired at the BS level (Daniel, 1969).

The PhD degree is available to those who wish to pursue more advanced training. Their goal may be teaching applied statistics, seeking a job in an industrial statistics research group, or pursuing a career with one of several high technology companies which need highly trained people to understand their processes and apply and develop appropriate statistical methodology. Obtaining a PhD degree will maximize one's opportunities as a statistician; however, it may not be the only strategy for other objectives. Movement to executive ranks and through production organizations may be equally well served with a MS degree.

MS and PhD statisticians often work on the same type of problems and use similar statistical techniques. Differences exist in the breadth and understanding of statistical methodology they can bring to bear on a problem, their ability to understand the technical problems involved, and the management levels with which they typically interact. Over a period of years many MS and some BS statisticians acquire by experience the necessary skills to function as a PhD statistician.

There is also the question of whether one should interrupt the formal education process to obtain industrial experience say between the MS and PhD degrees. While in theory this approach seems attractive, in practice, it extends the time span of one's formal education, and, in some instances, family and other responsibilities preempt the best intentions to return to school. Consulting intern programs, if properly organized, can give students the necessary practical experience without interrupting their training.

3.5 Part of Faculty and Student Research Should Flow from Real-World Problems

A subset of the faculty involved in training industrial statisticians should be engaged in statistics research which is stimulated by real-world problems and is directed at the improvement, classification, and extension of existing methodology which is of practical value. The statistical level of this work should be whatever is needed to develop useful solutions. The following, which is a paraphrase of TECHNOMETRICS editorial policy, reflects what we feel is an appropriate goal. Research projects shall include adequate justification of the application of the technique by means of numerical examples of application to a real-world problem in any subject matter area. The results of such studies can be published in a broad spectrum of journals such as TECHNOMETRICS, BIOMETRICS, APPLIED STATISTICS, and JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION. The publication of interpretative and expository articles in subject matter journals (eg, SCIENCE, JOURNAL OF QUALITY TECHNOLOGY, ARCHIVES OF ENVIRONMENTAL HEALTH, CLINICAL PHARMACOLOGY AND THERAPEUTICS) are also useful and should be encouraged and rewarded. These are particularly helpful in expanding the use of statistics in a scientific area as well as providing useful classroom examples.

PhD thesis research can be of several forms. It could be the solution of a real problem which requires a theoretical development. The solution, however, should hold up under careful scrutiny from a practical viewpoint as well as being sound theoretically. There are some companies which employ theoretical research groups and want to hire persons appropriately trained (admittedly, these represent a small fraction of the opportunities in industrial statistics). It may be appropriate for this work to be published in the ANNALS OF STATISTICS. In such an instance it may also be of value to publish an applications paper in a journal such as TECHNOMETRICS or BIOMETRICS, or a subject matter journal. Such a paper would illustrate how the new methodology would be used to solve the problem which stimulated the work.

Original research on a real problem (eg, involving innovative design of experiments and collection of data, data analysis, and model building) and the evaluation (eg, by analysis of many data sets) and extension of existing statistical methodology are two types of PhD thesis topics which could be considered for a student preparing for a career in industry. Such problems can be a natural by-product of the faculty's consulting activities. The former gives the student a "taste" for what his/her chosen career involves, and demonstrates an ability to formulate and solve a real problem. The latter demonstrates an ability to understand and advance the statistical theory and determine when and why it does and doesn't work. In each instance the thesis topic should, of course, be consistent with particular university's standards for the PhD degree. The resulting contributions should be original and make sense when viewed in terms of the practical realities associated with the problem. We recognize that this approach may depart from standard practice at many universities. Our intent is to suggest that such theses are worthy of consideration when preparing for a career in industry.

Research is less important at the MS level. In most instances this time is better spent working on real problems, gaining consulting experience, and developing communication techniques. A Master's degree project related to a real problem is usually of more value in this regard than a thesis (Kanji, 1979).

3.6 The Faculty's Interest in Real Problems Is Important

Many in industry feel that students' "upbringing" will carry over into their work and they weigh the philosophy of the faculty and university program quite heavily when hiring statisticians. Students' principal, and sometimes only, contacts with industrial statistics are the faculty members who are their instructors. It is important, therefore, that the faculty presents an accurate picture of the field of industrial statistics. The following discussion highlights faculty background, scientific interests, publications, and professional activities which are characteristic of many of those training industrial statisticians. We, of course, do not expect that, at a given university, all faculty will have these characteristics.

A portion of the faculty should have a keen in-

terest in solving real problems and orient their courses accordingly. The importance of problem solving can be emphasized by using publications (eg, BIOMETRICS, TECHNOMETRICS, J. QUALITY TECHNOLOGY) in course work which illustrate industrial uses of statistics. It is desirable to have had some full-time or part-time experience in industry. At least one degree and/or experience in an area other than statistics or mathematics is also desirable. Some of these experiences can be obtained through joint appointments, (although there is some disagreement over the effectiveness of these arrangements) and industrial consulting. Experienced faculty can also be obtained by appointing industrial statisticians as part-time coadjutant faculty. Having a portion of one's course work from a teacher who handles industrial problems daily can be very rewarding. The faculty should also have both broad scientific and statistics research interests because of the wide spectrum of problems encountered in industry.

We emphasize that statisticians consulting on real problems in a university should develop ideas and methodology worthy of publication in statistics journals. We also agree that subject matter publications are important, and should be rewarded. It is typical for the publication record of a faculty member working on real problems to contain several joint publications in the subject matter literature. A good consulting statistician will become "totally involved" in a project (Marquardt, 1979). The natural result of this involvement is an invitation to coauthor the resulting research report. This type of joint publication is indicative of an effective statistician and should be rewarded accordingly. The experience of coauthoring a paper with a nonstatistician will also improve the faculty members' communication techniques and give them examples of what is expected of their students. The faculty can relate these experiences to their students thereby improving the quality of instruction and demonstrating the relevance of statistical methodology.

Industrial statistics is a dynamic field and the faculty should keep up with the changes which are taking place. We believe that it is useful for faculty training statisticians for industry to take an active role (moderator, speaker, organizer, committee member) in both statistical and nonstatistical professional societies. This activity helps them identify what problems and concerns are important to the industrial statistician (eg, effect of Federal regulations on the practice of statistics in areas of health, environment, animal testing). Industrial statisticians attend the Gordon Research Conference on Statistics in Chemistry and Chemical Engineering, the American Society for Quality Control (ASQC) Annual Technical Conference, the ASQC Chemical Division - ASA Section on Physical and Engineering Sciences Fall Technical Conference, the Princeton Conference on Applied Statistics, and the ASA Annual Meeting, to name a few. It is also appropriate to encourage students to attend these conferences when possible.

3.7 The Faculty's Attitude Toward Statistics in Industry is Important

There are many different career alternatives for statisticians. Given accurate information on the alternatives, students will make decisions which are best for them. Some faculty members tend to want their best students to follow in their footsteps. This temptation should be resisted in favor of assisting the student to make the proper decision. As noted earlier, the growth of statistics in industry depends on the number of good statisticians selecting this career. To speed up the growth of this profession, all students should be made aware of the opportunities which exist. We will concentrate here on ways of bringing the field of industrial statistics to the attention of students.

An important piece of information concerning a career is the faculty's attitude towards it. Teachers' acknowledgment, through their teaching and research, that solving real problems is as important and as rewarding as publishing in prestigious theoretical journals will communicate to the student that industrial statistics is a worthwhile and reputable profession.

The importance of solving real problems can also be emphasized by introducing students to industrial statisticians at professional meetings and encouraging students to read articles on the activities of industrial statisticians. Currently there is no "careers brochure" aimed specifically at industrial statistics. It would be appropriate for the statistics' societies to initiate a project to fill this void.

4. WHAT CAN UNIVERSITIES AND INDUSTRY DO COOPERATIVELY TO DEVELOP APPROPRIATE PROGRAMS?

4.1 It Is Necessary to Understand Each Other's Needs and Problems

Improved communications between the university and industrial communities will speed up the development of appropriate programs for training industrial statisticians. These two groups must meet and interact often in a variety of forms in order to develop a mutual understanding of the issues. Some mechanisms which could be used to open lines of communication are:

- Visiting of universities by industrial statisticians and the visiting of industry by faculty members.
- Holding of "Career Days" by universities with invited industrial statisticians. This could be a joint venture of several neighboring universities.
- Development of industry internship programs.
- "Year in Industry" programs for faculty members and a "Year in the University" program for industrial statisticians.

Some of these suggestions may be difficult to implement; however, each should be considered to see if an opportunity for a fruitful exchange of ideas is possible.

When participating in these activities industrial statisticians should be careful to present their profession properly. There is a tendency on the part of some statisticians in industry to present papers which are academically oriented at professional meetings and university seminars. Meeting and seminar organizers and speakers should recognize the possibility of leaving false impressions and should plan carefully to ensure that the goal of the meeting and presentation is realized.

4.2 Developing Programs Which Meet the Needs of Industry and Universities

It is recognized that few departments can afford the luxury of a program designed specifically for industrial statistics. Rather the goal should be the training of practicing statisticians (ie, problem solvers) in general, with varying degrees of specialization in industrial statistics. A successful program will center around a faculty with the interests and characteristics discussed in Section 3.6. As noted in Section 1 it must also be determined what the needs of industry are, ie, what is the market?

A new program can be formulated by blending the approaches of existing successful programs with the perceived needs of industry and that of the given university. The needs of industry can be determined by

- Getting feedback from former students working in industry — most would be honored to comment.
- Soliciting of input from industrial groups.
- Participation of students and faculty in industrial intern programs.

A periodic review by industrial and academic statisticians is another possibility. This can be a delicate matter. The intent would be for industry to serve as a consultant to the university just as many university statisticians serve as consultants to industry. In the end the university is, of course, totally responsible for the content and emphasis of its program.

Industry can also help by bringing important statistical problems to the attention of universities. The success of this depends, in part, on the initiative of the universities. In general, if a problem is of keen interest to a company they will either solve it themselves or hire a consultant (often from a university) to develop the solution. Industry is unlikely to recognize an unsolved statistics research problem and support its solution unless they have persons with statistical training on the staff.

It is also reasonable for academic statisticians to contact industrial statisticians inquiring about important unsolved statistical problems and/or data sets (and problem backgrounds) which are appropriate for analysis by a specified statistical procedure. This approach will be most effective when both parties can benefit from the cooperative effort.

4.3 Continuing Education in the Practice of Statistics

There is some concern that there may not be enough experienced faculty to provide the type of training described here. Indeed, many university faculty members do not have the necessary practical experience. There are probably many industrial statisticians who would also be interested in improving their skills in consulting and the art of statistical practice. The holding of special conferences and training programs directed at the various aspects of industrial statistics, and the practice of statistics in general, is one potential mechanism for helping to fill this void. Some possible training programs might be

- One-day short course in conjunction with the ASA Annual Meeting
- One-week short course during the summer on a college campus
- Two- to four-week summer institute also held on a college campus
- One-year "postdoctoral type" program and its counterpart "year-in-industry" program.

We urge the profession to study this problem, create appropriate training programs, and explore methods of obtaining the necessary funding. A one- or two-day special conference may be the best way to get immediate and meaningful feedback.

5. CONCLUSIONS AND RECOMMENDATIONS

The conclusions and recommendations contained in this report are summarized and briefly discussed in this section. The background on each point can be found in the section referenced at the end of the associated paragraph. Our recommended programs focus on real problems and the statistical theory and methodology which are useful in their solution. As stated in Section 1, our purpose is to recommend approaches which will best prepare statistics students who wish to pursue careers in industry. Many of our views are consistent with the proposals of Greenfield (1979).

1. Programs for the training of industrial statisticians should contain a balance of theory and practice. The theory which is taught is that which is needed to understand the statistical methodology useful in practice. Relevant and useful theory is also important because industrial statisticians often have to adapt existing methodology to fit the problem and sometimes develop new methodology (Section 3.1).

2. It is important to include a wide variety of statistical techniques in the course work. Industrial statisticians are general practitioners and use many different statistical tools because of the wide range of problems and subject matter areas they encounter in their

work (Section 3.1).

3. Discussion of the realities of statistical practice is needed. It is important to discuss the realities of statistical practice such as problem formulation and analysis, importance of graphical displays, problems of historical data, and the use of the statistical literature. Students must understand how statistical methods are used in an industrial environment if they are to make significant contributions to the organizations which employ them (Section 3.1.1).

4. Students need hands-on experience with real problems. This is effectively accomplished by consulting intern programs. Working in consulting laboratories, taking consulting courses, and participating in joint seminars with other disciplines are also useful. Experience with real problems can also be included in course work; unfortunately, the face-to-face contact with the scientist is not possible in this situation (Section 3.2).

5. Enhancement of communication skills is a valuable addition to the program. The ability to communicate verbally and in writing with nonstatisticians at all technical and management levels is the key to the success of the industrial statistician. The preparation of written reports and associated critiques can be an integral part of the course work. Oral communications can be improved as part of the consulting experience, by oral presentations of statistical analyses (eg, consulting projects), and by teaching statistics courses to nonstatisticians (Section 3.3).

6. Real problems and associated data are essential in illustrating the statistical methodology being taught. Sufficient problem background should be given so the student can see how the statistical analysis contributed to the problem. These objectives can also be obtained, with varying degrees of success, by having students conduct their own experiments (Hunter, 1977) or critiquing and analyzing data reported in the subject matter literature. It is also appropriate to have a one or two semester course on "Analysis and Interpretation of Data" which is based totally on real problems (Section 3.2.2).

7. Students should be familiar with the well known statistical computer program packages and at least one programming language. An industrial statistician will likely have need for a computer early in his or her career. Statistics training programs should provide the necessary tools and experiences (Section 3.1).

8. Academic programs should be aimed at the MS degree level and higher. It is generally agreed that the MS degree is a minimum requirement for the professional statistician. There are always exceptions but someone interested in statistics is usually better served at the BS level to obtain solid foundations in a field of science or engineering, and mathematics by majoring in one of these fields and including a strong minor in the other. The PhD degree is available for those who wish to seek jobs that require more advanced training (Section 3.4).

9. Departments should offer opportunities for students' research which are stimulated by real problems. The statistical level is that which is needed to develop useful solutions. The results will be published in a broad spectrum of statistics journals with interpretative and expository articles being published in subject matter journals. PhD thesis research can be the solution of a real problem which requires a theoretical development. Original research on a scientific problem requiring innovative collection and analysis of data and the evaluation and extension of existing methodology through the analysis of many data sets are potential PhD thesis topics which are worthy of consideration (Section 3.5).

10. The faculty's interest in real problems is an important factor. It is generally felt that the student's "upbringing" carries over into his or her work and industry weighs the philosophy of the faculty quite heavily when hiring statisticians. A faculty's interest in real problems is evident in its course offerings, scientific interests, publications, and professional activities (Section 3.6).

11. The faculty's attitude toward statistics in industry is important. Students should be given accurate information on the various career alternatives so that they can make decisions which are best for them. Teachers of statistics can make the opportunities in industry known to their students through their teaching, research, and consulting activities. It is also worthwhile to introduce students to industrial statisticians when possible and to encourage students to read articles which describe the activities of industrial statistics and report the results of their work (Section 3.7).

12. A "Careers in Industrial Statistics" brochure is needed. The statistics' societies should initiate a project to fill this void (Section 3.7).

13. Universities and industry can work together to develop appropriate training programs. Improved communications are needed so each community can understand the other's needs and problems. Some suggested mechanisms are visits of universities by industrial statisticians, visits of industry by academic statisticians, the holding of "Career Days" at universities, the creation of industry consulting intern

programs, the appointing of industrial statisticians as coadjutant faculty members, and the arranging for faculty to spend a "year in industry". Suggestions on appropriate programs can be solicited from former students who are working in industry and from industrial groups (Sections 4.1 and 4.2).

14. The statistical societies and the universities should create programs to further address these issues. Two possible mechanisms are the holding of a special conference to discuss the "practice of statistics" and the development of special training programs and short courses for those who are interested in improving their skills in consulting and the art of statistical practice (Section 4.3).

ACKNOWLEDGEMENTS

This report was circulated to managers of statistics groups in a wide variety of companies for review and comment. Government statistics groups working in similar problem areas were also asked to participate in this review. Sixteen organizations representing the auto, chemical, communications, food, petroleum, pharmaceutical and tire industries and the government submitted comments. As expected, each organization identified one or two aspects which they felt should be emphasized more. They were in agreement, however, that the report accurately reflects the broad needs of industry. The input from these persons was very helpful and we are pleased to acknowledge their contributions. The committee also thanks D. S. Moore, Chair, ASA Section on Statistical Education, for his valuable guidance during the preparation of this report.

REFERENCES

Andrews, H. P. (1979). Personal communication.

Box, G.E.P., W. G. Hunter, and J. S. Hunter (1978). STATISTICS FOR EXPERIMENTERS, Wiley-Interscience, John Wiley and Sons, NY.

Daniel, C. (1969). Some general remarks on consulting in statistics. TECHNOMETRICS, 11, 241-246.

Greenfield, A. A. (1979). Statisticians in industrial research: The role and training of the industrial consultant. THE STATISTICIAN, 28, 71-82.

Hunter, W. G. (1977). Some ideas about teaching design of experiments, with 2^5 examples of experiments conducted by students. AMERICAN STATISTICIAN, 31, 12-17.

Kanji, G. K. (1979). The role of projects in statistical education. THE STATISTICIAN, 28, 19-27.

Marquardt, D. W. (1979). Statistical consulting in industry. THE AMERICAN STATISTICIAN, to appear.

Ott, E. R. (1975). PROCESS QUALITY CONTROL, TROUBLESHOOTING AND INTERPRETATION OF DATA, McGraw-Hill, NY.

Woodward, W. A. and W. R. Schucany (1977). Bibliography for statistical consulting. BIOMETRICS, 33, 564-565.

DISCUSSION

Paul D. Minton, Virginia Commonwealth University

The Committee is due much appreciation for an excellent report and for the thought, work and experience that it represents.

I am sure many of us have pet ideas on this subject. Mine differ little from the Committee's, but I would like to mention a possible addition or two that I think important, some suggestions on emphasis or amplification, and some aspects from an administrator's viewpoint, as I was requested to do.

One addition concerns work in industry for a Bachelor's degree holder. The Report discusses training for a professional statistician in industry, and defines such a person to be one trained at the Master's level or higher. My limited observation suggests that this is correct for those organizations which have begun to use the professional consulting statistician or groups of them. However, I believe that there are many more industrial organizations that have not than have, and that these are engaged in a considerable amount of statistical activity at the Bachelor's degree level, carried on by personnel with degrees in everything but statistics. One could argue that that is proper, but I do not think so. One could argue that the company should hire a full time Ph.D. statistician , but the company's perception is different. For such companies, a bachelor's-level statistician would do a better job than those now getting these statistical assignments, and would be able to recognize better when the problem is beyond his or her level of training and when the volume and level of statistical activity indicate the need for a full time professional statistician.

Some of these employees would continue into graduate work, part-time perhaps, some with company aid. These would need to study some mathematics that they could have studied as undergraduates, as the Report suggests, but on the other hand they will not need the industry-oriented segment of the graduate program, and will have an advantage in perspective and interest.

The Report has laid out quite well for us the Master's and Ph.D. level of training in statistical theory and methodology needed by a statistical consultant in industry. I would like to see us push our Bachelor graduates for jobs that <u>are</u> suitable to their training, just as many other disciplines do. Thus, I would like to see a section in the report, or a separate report, on suitable jobs for baccalaureate degree holders, and would suggest separate tables of characteristics desired, and services graduates could be expected to perform, for each of the three degrees.

Another addition concerns industrial management's view of statistics. Except in a relatively few enlightened companies, statistics is not as well accepted as a necessity as it should be (see Professor Harry V. Roberts' article "Statisticians Can Matter" in <u>The American Statistician</u>, May, 1978.). The Report makes a good point that the industrial statistician's successful practice will do much to sell statistics in industry, and that part of his job is to do so, but the discipline as a whole has a responsibility for selling,also, and we have done little to tell our story to industry.

The Report recommends a booklet for recruiting people into careers as industrial statisticians. I would like to see, in addition, a booklet advertising the value of statisticians, at all degree levels, to industry. Such a booklet could lay out what persons at each degree level could be expected to do and how that would be profitable. The booklet would be aimed primarily at companies without professional practicing statisticians, although it would also help the struggling young practitioner just breaking into a company all alone. He must sell statistics in addition to the big job of learning to apply statistics well in his company.

I have a few comments on wording or emphasis. Regarding the profile of an industrial statistician, I would note that backgrounds include not only science, engineering and mathematics, as mentioned in the Report, but also social sciences, business systems, information sciences and others. Also, many of the characteristics of an industrial statistician listed in the table seem to fall under the general heading of "breadth", and I would like to see this recognized and emphasized in the text, including knowledge of the subject matter of the client, willingness to study that subject matter and to research the statistical literature and methodology already adapted to that field, in addition to ability to devise new methodology.

Finally, I have some comments from an administrative point of view. Courses are the province of the faculty, so that one problem in proposing course changes is to achieve some consensus in organizing topics and courses to achieve a reasonable sequential presentation from the students' standpoint while maintaining sufficient enrollment in each course to justify regular offerings. The Report's list of statistical techniques used most often by industrial statisticians includes mostly topics that would be included automatically in one form or another in most Master's or PH.D. programs, so that only a few would present any problem. The suggestion of mini-courses to make students aware of the statistical literature in fields not studied in detail seems to me to be a good one, although not necessarily easy to implement.

Courses on consulting techniques and effective communication should be undertaken jointly with faculties of psychology, English, or other appropriate disciplines; as statisticians in perhaps the field most poached upon by other disciplines, we must be above the suspicion of becoming poach-

ers rather than poachees. It seems to me to be best for the student to have communication taught by experts in communication. When there is a disagreement over who is to teach what, it sometimes becomes an administrative problem, but one that an administrative officer has only a limited power to solve, since it involves the province of the faculty. It is clear to me that when there is lack of cooperation, both sides lose; it would be just as easy for both sides to win, by simply agreeing to cooperate in order to achieve that mutual profit.

Of course, in addition to communication aspects, there are aspects to consulting that are strictly statistical that we must get across to students. See, for example, Allyn Kimball's 1959 paper in JASA on "Errors of the Third Kind in Statistical Consulting" - giving the right answer to the wrong problem!

As the Report notes, communication skills can be enhanced by all parts of the program. Experience at several universities with various means of encouraging graduate students to study together indicates that this is one of the best ways of increasing communication skills; these students were "wrung out" more by each other than by anything else. Allocating and managing space for such purposes can present administrative problems.

Hands-on experience with real problems is essential, as the Report notes. My experience is that there are administrative problems to be faced if this is to be done well for all students. An internship course carries problems of credit for the students and for the professor, possible payment for the student, assignments of space, status of student as employee or non-employee, keeping of office hours, facing deadlines, provision for computer service, etc. If there is not an ongoing program of consultative assistance in research by faculty, the problem of supply of projects must be faced. Involving students in outside consulting as a part of required training would be difficult. Friends in industry have found it difficult to supply illustrative problems for students. In sum, however, though not easy, experience with real problems is very much worth the effort, as the Report says.

Joint seminars with other disciplines in academe, which you mention, are also worth the effort, although they represent again the problem of relationships among disciplines. These relationships are not uniformly cooperative ones, and the only way for statistics groups to improve our relationships with other disciplines is for us to take the lead by reaching out our hands giving away free samples, etc., if we are to play a significant role in purveying statistical instruction and consultative assistance campuswide. Where there is cooperation, the administrative role is one of assistance; where there is not, cooperation is seldom produced by administrative coercion.

A MODULAR SELF-PACED STATISTICS COURSE: FIVE YEARS' EXPERIENCE

Donald A. Anderson and William P. Cooke
University of Wyoming

1. Traditional Course and Motivation for Change

The Department of Statistics at the University of Wyoming is part of the College of Commerce and Industry, with about 40 per cent of the students enrolled in its junior-senior level service courses coming from the Departments of Business Administration, Accounting and Economics. The beginning course, however, is a first statistics course for students from all colleges in the University. It usually has an enrollment of around 300 students per semester, with approximately a third each from the general areas of Social and Applied Social Sciences, Biological Sciences, and Physical Sciences and Engineering. For example, in one recent semester 45 per cent of the enrollment consisted of students from the College of Arts and Sciences, majoring in 19 different areas, 38 per cent came from Commerce and Industry, representing seven different major areas, and the remaining 17 per cent consisted of majors in 15 departments from the Colleges of Agriculture, Education, Engineering, and Health Sciences. These percentages are generally representative of enrollments in any semester.

If students from other colleges must take service-type, research methods courses beyond this beginning, minimal-prerequisite introductory course, they usually take them from the Department of Statistics. However, some departments, such as the Departments of Sociology, Psychology, Zoology, Geology, and Educational Administration, offer their own specialized research methods courses for which our first course is a prerequisite. Thus we must be sure that people receiving credit for our beginning course are adequately prepared for our own courses in sampling, regression, analysis of variance, and nonparametric statistics, as well as for the specialized courses offered by others.

Until the 1974-75 academic year, introductory statistics had been taught by the traditionally accepted formula of lecture-laboratory-test. That four semester-hour course involved three lecture hours and two laboratory hours per week. Lectures were presented by the faculty to two large sections of approximately 150 students each, while the problem-solving laboratory sessions, limited in size to 30 students, were handled by graduate assistants majoring in statistics. A student would normally expect his course grade to be based on performance on three or four hour exams, short lab tests, and a final examination.

Continually, the faculty of the Department of Statistics had two main areas of concern regarding the basic course. First, historically only 50 to 60 per cent of the students enrolling in a given semester would successfully complete the course. The remainder would either withdraw early, withdraw later with an F grade, or struggle through the entire course as borderline cases, eventually earning a grade of F. Secondly, many students would bring to the course an apprehensive or resentful attitude. They had heard that statistics was difficult, and had from the outset resigned themselves to an F, a squeaky D, or perhaps a lucky C. Often this attitude was a result of a poor high school mathematics background, or because they had deferred the statistics course until their senior year and had forgotten what little algebra they had learned in high school. But often this bad attitude could be traced to the student grapevine, which reliable source had labeled our course an overly difficult one to pass.

In the spring of 1974 the department, at the urging of Professor Lyman McDonald, decided to try changing student attitude and success rate by completely redesigning the mode of presentation of the material. Throughout the summer of 1974 Dr. McDonald and two graduate assistants prepared units of instruction and large sets of problems for each unit, and assembled a Student Guide for the modular, self-paced course to be described in Section 2. The new version of the course was initiated in the fall of 1974, along with one traditional section so that comparisons could be made.

It should be noted that a distinct change in student attitude was perceived from the moment the self-paced course began. The students were aware of the substantial changes and the "trial basis" nature of the class, and appreciated the fact that we were concerned enough about their difficulties to attempt to remedy deficiencies in the course. Thus one of our objectives had been partially achieved merely by our decision to experiment. It should also be pointed out that now, five years later, with the self-paced course no longer an experiment but a fixture, this advantage has disappeared. We must now depend upon the quality of the course and its instructors to achieve the desired good student attitude.

2. The Original Modular Self-Paced Course

William C. Guenther's book, Introduction to Statistical Inference, 2nd ed. (McGraw-Hill, New York, 1973) was selected as the basic text. The material to be covered was decided upon and was organized into 11 units. For each unit a set of objectives was prepared to permit the student to isolate the important points in the unit and to locate appropriate reference material.

A clear statement of objectives for each unit is an essential part of the course. By organizing the units and objectives, the instructor is compelled to isolate the essential content of the course and to fairly report it to the student. A portion of the list of objectives for Unit 2 is included here as Table 1.

TABLE 1. SAMPLE UNIT 2 OBJECTIVES, RANDOM VARIABLES AND PROBABILITY MODELS

36*	1.	Be able to define "random variable."
36-40	2.	Given a probability distribution table for the random variable X, be able to a. Evaluate $Pr(X = x)$. b. Construct a cumulative probability distribution table. c. Evaluate $Pr(X \leq x)$ and $Pr(X \geq x)$.
	⋮	
49	5.	Be able to define "continuous random variable" and "discrete random variable."
48-52	6.	Given the density function, $f(x)$, for a continuous random variable X, give the two basic properties of $f(x)$.
37		Understand the notations: $f(x)$ and $F(x)$.
	⋮	
		Binomial Probability Distribution
58-59	9.	Given an experiment where two possible outcomes called "success" and "failure" have been defined and where the probability of success is p, calculate the probability of failure, q.
58-65	10.	Recognize when the binomial model is applicable.
58	11.	List the four conditions necessary for the application of a binomial probability model.
	⋮	

*Page number from text.

The second step in organizing the course was the writing of a large number of exercises to be used as test items and laboratory problems for each unit. Approximately 60 exercises for each unit were prepared and, when appropriate, were written in language suitable for each of the three areas: Social Sciences, Biological Sciences, and Business. This enabled students to see and work with applications generally in their area of interest.

The third step was the staffing and equipping of rooms for (1) an "open lab," (2) an "open testing room," and (3) lectures. The open lab and the open testing room were equipped with tables and about 30 electronic calculators each, with the testing room also having locking filing cabinets for test storage, extra calculators, and other supplies. The rooms were reserved for nearly seven hours per day, Monday through Friday. In addition, three lecture rooms were reserved for the same hour, three days a week.

Staff for the open lab consisted of four graduate assistants and 15 undergraduate tutors. At every hour the lab was open one graduate assistant and three or four undergraduate tutors were available to provide individual help to students working on laboratory exercises. The undergraduate tutors were students who had recently finished the course with an A grade. They received one semester-hour credit for three hours per week under a special problems course number.

The open testing room was manned by two graduate assistants, who administered and graded tests on the units. Further, in addition to their lecturing duties, each of the three faculty members assigned to the course worked about four hours per week in the lab and testing rooms.

By a random process the exercises for a given unit were split into three "laboratory exercises" (one each for Social Sciences, Biological Sciences and Business) and a "pool of test questions." Stratified random sampling from that pool constructed the tests. To satisfactorily complete a unit the student was required (1) to complete a laboratory exercise and have it checked in the open lab, and (2) to pass an individual examination (about eight questions from the test question pool) with a grade of at least 90 per cent in the open testing room.

Exams were graded immediately, with the student present. Students could call for additional examinations without penalty until a passing score of 90 or better was achieved, but were restricted to one attempt per day.

The units were organized so that the first eight units more or less covered (but not at the same depth) the traditional course that had been taught in previous years. The ninth and tenth units recovered some of those topics in more depth and picked up other special topics often included, at the instructor's discretion, in the traditional course. The eleventh and final unit was an introduction to nonparametric statistics, a topic not usually included in the course prior to 1974. Topical coverage of the course is given in Section 4.

The basic requirement for a final grade of C was satisfactory completion of the first eight units, for a B the first 10 units, and completion of all 11 units was required for an A. Additionally, the students were required to (1) pass a final examination over the appropriate units, with the minimum acceptable score being 60 per cent, (2) participate in at least one special project that had been approved by the instructors (e.g., help conduct an opinion poll and analyze the results), and (3) attend at least three guest lectures. The guest lecturers were faculty

members recruited from departments around the campus who were asked to speak about uses they have made of statistics in their research.

Regular course lectures were delivered by the instructors in different rooms but at the same hour so students could freely transfer from one lecturer to another. One instructor paced his lectures to complete the first eight (C) units by the end of the semester. Similarly, another covered the first 10 (B) units, and the third covered all 11 (A) units by the end of the semester. Attendance at lectures was not required.

To minimize misunderstandings of the content and conduct of the course, each student received a mimeographed Student Guide to the course. This guide carefully laid out all of the requirements for successful completion of the course and requirements for grades of A, B, C (and D and F), included all objectives for all the units, and contained advice on good study habits and additional references for study outside of the basic textbook.

3. Evaluation of Initial Year

At the end of the first two semesters' experience with the modular self-paced course, the three instructors were asked for their candid opinions by the Dean of the College of Commerce and Industry. Further, students in each semester responded to a questionnaire requesting their viewpoint about various facets of the course they had just completed. These student and faculty evaluations were reported in some detail in a paper entitled "The University of Wyoming 1974-75 Modular, Self-Paced Introductory Course in Statistics," College of Commerce and Industry Research Paper No. 97, by Lyman L. McDonald, Timon A. Walther, and William P. Cooke. Only those parts of that evaluation bearing on changes that have been made in the course since then will be reported here, with the exception of the following two paragraphs, which indicate student acceptance of the procedure.

The hoped-for improvement in student attitude was definitely achieved. When asked whether they would advise a friend to enroll for the self-paced version instead of a traditional section, 23 per cent responded, "Yes, with no reservations," while another 60 per cent said, "Yes, with some qualifications." The principal qualifications listed were (1) the need for self discipline, (2) the need of a good background in algebra, and (3) the need to allow adequate time to complete the work.

Eighty-one per cent felt that more self discipline was required to succeed in this course than with the usual kind of class to which they were accustomed. Also, 70 per cent of the students responded with the (slightly surprising) comment that they felt they received more individual attention in our course. Another 61 per cent said they very much enjoyed having the freedom to plan their own times for study and laboratory sessions, and to call for examinations at their convenience. Merging these responses with the additional one that 82 per cent felt that they would remember the material for the same length of time or longer than they would have under the old, rigid style, of course the three instructors concluded that --from the students' viewpoint at least--we had succeeded in presenting a better beginning statistics course.

Some aspects of the course, however, from the staff viewpoint, were a bureaucratic nightmare. Because of the necessity for students to satisfactorily complete eight units by the end of the semester, in the fall of 1974 there were eight firm "deadlines." Students had to complete units according to that schedule or else withdraw from the course if a deadline was missed. This created havoc in the testing room precisely eight times during the semester, and a withdrawal rate of 37 per cent, hardly an improvement in the success rate. So, in the spring of 1975 we dropped the required completion schedule and instigated a reward system for early completion of units (advance points toward the final examination). This merely deferred the congestion in the testing room to the last two weeks of the semester, when some students were frantically trying to pass as many as four units, and the withdrawal rate for the course jumped to 47 per cent. Somehow, the seemingly good idea of complete student freedom to choose the time of their unit exams and the essentially "unlimited retakes" feature produced too often an unmanageable situation (due to lack of personnel) in the testing room.

Not requiring attendance at lectures resulted in about 50 per cent of the students attending no lectures after the first two weeks of the course (this percentage includes those who withdrew early or intended to withdraw later on). From their point of view, the objective was to work just enough laboratory exercises to permit them to score well on the unit exams--not necessarily the same thing as gaining a good appreciation of the value of statistical methods to support or cast doubt upon the validity of scientific hypotheses. We had hoped that what we had to say would add to the students' appreciation of statistics and to their understanding of its philosophical basis. Those students who did attend lectures regularly (and who became predominately the B and A-graded group), did indicate that they thought the lectures were worth their having attended. But fully half of the students enrolled had not come to listen to them.

This lack of attendance at lectures caused an unforseen problem in the open laboratory. Instead of being able to move rapidly from student to student, responding to intelligent questions about laboratory exercises, the student assistants found themselves being asked essentially to give individual lectures to one or two students at a time. This feature drastically reduced the efficiency we had envisioned in the open labs.

The new method, while not improving the overall percentage of students receiving credit for the course, did produce a change in grade distribution. The percentage of A grades nearly doubled, but this result seemed to be due to students who might have been graded B under the old

TABLE 2. GRADE DISTRIBUTIONS (percentages) OVER FOUR SEMESTERS

Grade	Lecture-Laboratory-Test System			Modular Self-Paced System	
	Fall 1972	Spring 1973	Fall 1974	Fall 1974	Spring 1975
A	16	12	19	27	29
B	25	22	13	9	2
C	27	29	17	22	15
D	12	8	7	2	0
F	4	7	22	3	7
Withdrew	16	22	22	37	47

system being motivated to extra effort. Comparative grade distributions are shown in Table 2.

Student response to questionnaires also revealed that (1) the passing score of 90 per cent on unit exams was regarded as unreasonable, (2) many students felt that grading of exams was inconsistent from grader to grader, (3) the special project requirement was very popular, and (4) the guest lectures were the least appreciated facet of the course. All of these observations, along with consideration of improved student attitude, high withdrawal rate, and the new grade distribution, were taken into account as the department undertook an end-of-year evaluation of the course.

The decision was to continue the course under the modular system, but to cut back on student freedom and to instigate better control procedures. We still liked the idea that students had to demonstrate competence on one unit of material before proceeding to the next, and we were impressed by the improved student attitude even in the face of our failure to improve the success rate. The Dean was most impressed by the fact that hardly any students had run to his office with a complaint, a problem all too prevalent in the past. Students, whether those earning an A or those who had withdrawn, could easily understand that regardless of some of our fumbling administrative procedures, the responsibility for success or failure was basically theirs.

4. Evolution over the Next Four Years

It was clear from the first year's experience that three basic changes were absolutely necessary. They were (1) required attendance at lectures, (2) a system of firm deadlines for unit completion, and (3) a limitation on the number of tries on a unit exam before withdrawal from the course became mandatory. Requiring attendance at lectures would assure that students had at least heard about the type of problems they would encounter in the open labs, and would promote greater student-instructor contact. The other two changes would make efficient operation of the testing room feasible with limited personnel.

Only two lectures, an "A-pace" and a "B-pace" lecture, are now run for the entire semester. Hardly any student now resigns himself to a C from the outset of the course. Provision is made for starting up a slower-paced third lecture about mid-semester for those students who have fallen behind the pace of the B lectures.

The kind-hearted segment of the department won the point that the score for satisfactory completion of a unit exam should be reduced from 90 per cent to some lower figure. The number of attempts permitted on a single unit was set at three, and with this limitation it was felt that a lower passing score should be allowed. Depending upon which professor has had primary responsibility for the course, this score has been at various times over the past four years, 70 per cent, 80 per cent, and 85 per cent. This magic number has now settled down to a constant 80, always accompanied by the "three tries and out" rule.

There has also been some experimentation with the content and the number of units. The basic course content has remained essentially the same, with the A grade given for mastery of a bit more material than was found in our traditional course. Curiously perhaps, the main fatality has been the special project. While students had enjoyed this aspect of the course (it made them think they were doing real statistics), the special project was unpopular with the faculty. One reason was the large block of faculty time necessary for approving, directing, and evaluating the projects. However, we like to think that the primary reason for its abandonment was that most of these projects were small and trivial in nature, not really demonstrating the actual conduct of a major research study. During one recent semester, a "Data Collection Unit" was added in addition to the now standard number of 10 basic units. This turned out to be a feeble attempt to replace a hands-on project with lectures about real projects and was roundly criticized by the students as a weak unit of work.

Changes have also occurred in the handling of the final examination and in the requirements for grades. Initially, a score of 60 per cent or better was required on the final if a student was to keep the grade he/she expected at the end of the year. If less than 60 per cent was achieved on the final, the grade was reduced one letter. In the second year, the passing score on the final was changed to 60 per cent for a C, 70 for a B, and 80 for a grade of A, still in conjunction with the reduction in grade for failure to demonstrate those levels of competence. In the meantime, a "point total" system had also been added, so that not only must a minimum number of units be satisfactorily completed, but a prescribed number of points must be accumulated to earn the desired grade. This combination of requirements resulted in some students with lower point totals receiving higher course grades than

others who had not done well on the final exam. This element of apparent unfairness has since been removed.

Now only fixed point totals are required for the grades of A, B, C and D, with the passing score on unit exams fixed at 80 per cent. No specified numbers of units are involved. For a grade of B or C, for example, a student may either complete enough units to achieve the published number of points, or may earn his last few points by making a good score on the final examination, which is no longer required, but optional. The total number of units is now 10, and point totals are established so that, roughly, an average of 85 points per unit on 9 and 10 units will produce a C or B grade, while an 85 average on all 10 units and approximately 70 on a final (necessary for an A) will yield an A. This has proved much more popular with students and faculty, and has the added advantage that most of the students receiving credit for the course will have studied the material in all 10 units.

A drastic change in open lab procedure has been recently instigated. Firm deadlines for passing unit exams are now scheduled that each deadline may be preceded by two laboratory sessions, one of which is devoted to problem solving with tutoring available, with the other session involving a practice examination. This almost sounds like a retreat to the traditional format, and in a way it is, except that the laboratory is still completely open every afternoon for working on problems with help if desired.

Another retreat from full-bore self-pacing has been the way testing is handled. Now there are regularly scheduled examination days in the laboratories. Students can no longer call for a unit exam at any time, but must take the unit exam on whatever unit they are currently studying at a specified time. This takes away a lot of student freedom, but has made the testing an easily manageable part of the course. There is enough flexibility in the schedule to permit as many as three tries on all but the last (10th) unit to occur on scheduled exam days. Self-pacing remains, however, in the sense that students still have some choice about their rate of completion of units and the number of units they choose to complete.

Most of the changes obviously take away student responsibility to gain administrative tractability. Student morale has remained high, however, and the grade distribution has now changed drastically . Required lectures, enforced deadlines, and streamlined administration has produced a high percentage of A and B grades and a reduction in withdrawal rate. Most of these gains, however, have occurred during the past year. Table 3 exhibits some of the evolution of the grade distribution.

Note the similarities of grade distributions during the earlier semester of experimentation with the course, and the decidedly different distribution in the most recently completed semester. We feel that this pleasing result was due to the current method of handling laboratories more formally and the restriction of examinations to specified days (changes initiated after Spring, 1977). Limitations on timing of examinations has made it a bit more difficult for the average student to earn an A for perseyerance alone, although this one-time C student may now be in our B category. Required lecture attendance along with the more structured character of the course is, we think, the reason for the substantially reduced withdrawal rate.

The discussion to this point has been on structure, administration, and results achieved with the modular self-paced beginning statistics course. But what about its content?

We believe that the content of our four-semester-hour course is quite substantial, and that the level of performance required of students is high. Table 4 gives an annotated list of unit topics.

5. Evaluation of Current Course

The two main objectives of five years ago were achieved. There has been a definite improvement in student attitude, and the success rate has been increased substantially. Students with little mathematical or scientific background still find beginning statistics difficult, but the specific objectives and directed study features of the course are allowing them to succeed.

The performance level of the average student during the course has been increased, along with an increase of topical coverage over that in the traditional course. Students must demonstrate competence on preceding units of work before going on to other topics.

There has been some discussion within the department, with no clear consensus yet, about two disturbing aspects of the procedure. The percentages of A's and B's are running high, although not too high relative to current national trends. Still, we wonder whether an A in our beginning course means the same thing as an A in one of our advanced undergraduate methods courses. Is the A given for outstanding performance or only for peseverance? Secondly, there has been little evidence in courses for which the beginning course is prerequisite that the students are any better prepared upon entry. We agree that performance in second-level courses is no worse than before, but cannot detect that it is any better. In fact, quite often the reversion to a traditional method in a second-level course has surprised and disconcerted students fresh from the modular self-paced version. They much prefer the method used in their first course.

In sum, we have succeeded at presenting a beginning course in statistics which, if not enjoyable, is at least palatable to our students. We feel that the improved student success and attitude have certainly made the effort worthwhile.

TABLE 3. GRADE DISTRIBUTIONS (percentages) FOR FOUR MORE SEMESTERS

Grade	Fall 1975	Fall 1976	Spring 1977	Spring 1979
A	47	31	49	34
B	9	7	6	33
C	14	12	12	18
D	0	0	1	2
F	6	5	4	5
Withdrew	24	45	28	8

TABLE 4. COURSE CONTENT

Unit	Topic
1	Elementary Probability -- sample spaces, events, addition and multiplication formulas, counting formulas.
2	Random Variables and Probability Models -- properties of random variables and probability distributions, binomial and hypergeometric models.
3	Standard Normal Probability Distribution -- properties of standard normal curve, use of tables, sampling distribution of $\bar{X}$, statistics, estimation, central limit theorem.
4	Hypothesis Testing and Confidence Intervals -- one and two-tailed tests, errors, power, tests and confidence intervals for mean with variance known.
5	Hypothesis Testing and Estimation -- t-distribution, tables, one and two-sample t-tests, paired t-tests.
6	Testing and Estimation on a Proportion p -- one and two-sample tests and confidence intervals for one and two binomial proportions and differences in proportions, the chi-square distribution and tables.
7	Approximate Chi-Square Tests -- multinomial proportions, contingency tables, homogeneity and independence hypotheses.
8	Correlation and Regression -- simple linear regression formulas, confidence intervals and tests for slope, prediction intervals, correlation and relationship of correlation problem to regression problem.
9	Distribution-Free Inferences -- quantiles, sign test, Wilcoxon signed rank and rank sum tests, Spearman rank correlation.
10	Analysis of Variance -- one-way classification ANOVA, assumptions and computational formulas, dot notation, contrasts, Scheffé intervals.

STUDENT'S CHOICE -- SELF-PACING AND/OR THE LECTURE METHOD FOR INTRODUCTORY STATISTICS

by Howard B. Christensen, Brigham Young University

INTRODUCTION

When an introductory statistics course is discussed among a group of instructors, one or both of two types of courses may be implied. One type of introductory course might be the so-called "statistics appreciation" course, a course whose main objective is not to present statistical methods as such, but is intended to present the non-mathematical student with some underlying statistical concepts. Typical of these topics are concepts of proper experimental design, concepts and procedures of probability sampling, the effect of variability, the purpose of randomness, some basic descriptive procedures, etc. These courses assume very little or no math, with computations no more complex than that of computing a mean. Several texts in this area are: Statistics by Freedman, Pisani and Purvis; Statistics, A New Approach by Wallis and Roberts; Statistics, Concepts and Controversies by Moore, etc. The Statistics: A Guide to the Unknown series represents a good set of companion volumes for such a course.

The other type of introductory course is the more traditional course whose topics might be chosen from the techniques of descriptive statistics, probability, sampling distributions, confidence intervals, t-tests, χ^2 tests for contingency tables, linear regression, introduction to analysis of variance, and so forth. This type of course requires more mathematical background with the texts by Huntsberger, Freund, Mendenhall, and others being typical texts for such a course.

At Brigham Young University we have both types of courses. In both cases they are primarily service courses to other departments in the university community. The topic of this paper deals with a description of our approach to the latter, more traditional introductory course in statistical methods.

HISTORICAL BACKGROUND

From the time the Department of Statistics was formed in the early 60's up through about 1972, our course was taught by the lecture method in sections of about 30-40 students each. The clientele for this course came mainly from the College of Business. However, more and more departments were requiring their students to take a basic statistics course and thus the total number of students we served showed a steady increase over that time span. To handle this increase in enrollment, we ended up assigning more and more of our sections to be taught by graduate students, many of whom proved extremely capable in this assignment. However, we had several concerns that finally brought us to a change of direction in our approach to the course. First, we were concerned about the consistency of topics covered and the quality of presentation from section to section as more and more sections were taught by more and more different individuals. Secondly, this course represented an important source of undergraduate majors, and we felt that the source was in danger as fewer of the courses were taught by full-time statistics faculty who could do some recruiting. Thirdly, we had long been aware of student disenchantment for the course in terms of its relevance to specialized areas or majors; we were also aware that we were developing a reputation of having one of the more difficult courses in a student's curriculum. In fact, evidence tended to indicate students were putting off a Sophomore-level course until their Junior or Senior years, and then taking the course with a certain amount of fear and worry.

Through a variety of actions taken in 1972, a working arrangement was set up between the Department of Statistics and the Department of Instructional Development (DID) now a part of the David O. McKay Institute of Education. The objective of this cooperative effort was to develop instructional strategies intended to correct the three concerns we had as a department.

A preliminary study was undertaken to validate or invalidate our preconceived ideas about student reaction to the course. The study identified seven problem areas (Green and Christensen 1977):

1. Other departments had not felt that the Statistics Department was making its elementary statistics course relevant to the peculiar problems commonly met by majors in sociology, psychology, education, physical education or business.
2. Practical application of most statistical concepts was not apparent to the students, causing them to be bored and negative about the course.
3. The reputation of the course was such that the students delayed taking it until their junior and senior years, although it was orginally conceived as a sophomore-level course.
4. So many different instructors were teaching the course that great variability existed among the sections of the course being taught.
5. Many lecture sections were not being taught by full-time faculty members, but by inexperienced graduate assistants.
6. Lab space was inadequate.
7. The lab manual was not coordinated with problems in the lecture, the text, or the examinations.

To address these seven problem areas a senior author from the Department of Statistics was assigned to work with the personnel from DID to develop course materials, examples, tests, etc. to fit into a framework designed by DID to improve our course.

One of the first steps was to meet with representatives of all other departments who had students taking an introductory statistics course. Our purpose was to define what topics they expected their students to get from an introductory course. Compromise was an essential ingredient in these meetings, but out of it came a course description that all participants recognized as practical under the circumstances.

The next step was to develop, under direction and guidance of DID, mappings that showed the interrelationships of concepts associated with the course description defined by the interdepartmental committee. This was no easy exercise and a great deal of effort went into it.

At the conclusion of this phase, DID took the information generated and proposed a strategy for teaching the course. (Up to this point there had been no prior assumption made as to the media that would be used for instruction--films, slides, tapes, books, etc. were all possibilities--the final form had intentionally been left undefined so that as few constraints as possible were imposed in the earlier, developmental stages.)

The proposal from DID was to prepare a textbook using concepts they had developed for use on a computer assisted instructional program called the TICCIT project. This textbook was to be self-pacing but was not a programmed text. In addition, a large section format was proposed with an instructional management system devised to monitor the student's progress through the course.

This system was implemented on an experimental basis the Fall of 1973. It was implemented on a full scale in the Fall of 1975 and has continued to the present with both major and minor revisions over that time span. The basic components of that system will be discussed in the following sections.

INSTRUCTIONAL MATERIAL

The Textbook:

The textbook (Christensen, 1977) consists of Units (chapters) and Lessons within Units. Each Unit consists of a broad class of related concepts, e.g. one units covers graphical methods, another covers confidence intervals. A given lesson in a unit covers a single concept or a small group of closely related concepts. For example, one lesson in the Unit on confidence intervals covers the concept of unbiased and efficient estimators, another the concept of a standard error, still another lesson covers the concept of a confidence interval for a single mean, etc. The distinctive feature which makes the text different from other texts is the consistency of format from lesson to lesson.

Each lesson is made up of the following parts:

1. The Task Statement: A statement is made of the skill or concept to be learned and applied in the lesson.
2. The Definition: A concise definition of the concept or concepts is given in a short definition section.
3. The Discussion Section: This is an elaboration on the definition(s) with simple examples to put the definitions in a framework that the background experiences of the student would help him/her to understand better. Interrelationships of concepts would be pointed out.
4. The Simplified Discussion (Optional): In those situations where the experiences of the author indicated the concepts under discussion were particularly difficult to grasp by the student, an extra, simpler discussion using analogies and detailed examples was included.
5. Example Problems: Examples with detailed explanations are given to place the concept in concrete terms for application.
6. Practice Problems: Problems similar to the example problems are given to allow the students to test his/her skills. Answers to the odd-numbered practice problems are available to the student at the end ofeach Unit and the even numbered answers are in the teachers supplement.

Each lesson is set up using the same format which makes it possible for the student to work more efficiently through the materials as well as allowing him/her to develop a more personalized study strategy as familiarity with format and material increases. For example, one student may find that the definition and example problems are sufficient to understand the concept. Another might choose to go to example problems first and then back to definitions and then to practice problems.

To unify the lessons, they are grouped together in a logical manner, each unit containing a:

1. Unit Map which serves the purpose of showing graphically a logical "series" or "parallel" connection to the individual lessons in the Unit.
2. Introduction and Vocabulary section to point the direction the unit takes as well as identifying the new terms to be covered in the Unit.
3. Lesson Series, the basic pedogogical component in the Unit.

4. Application Section to describe how the material in the unit relates to real world settings as well as its connection to preceding and succeeding Units.
5. Summary Section to pull together in a capsulization, the various concepts presented in the Unit.
6. Unit Self-Test, a test that prepares the student to take a format test on the material--answers are provided at the back of the textbook.
7. Answers to Practice Problems, to provide immediate feedback to students on their mastery of the concepts and skills in each of the lessons.

The interrelationship of all these components are shown in Figure 1 that follows.

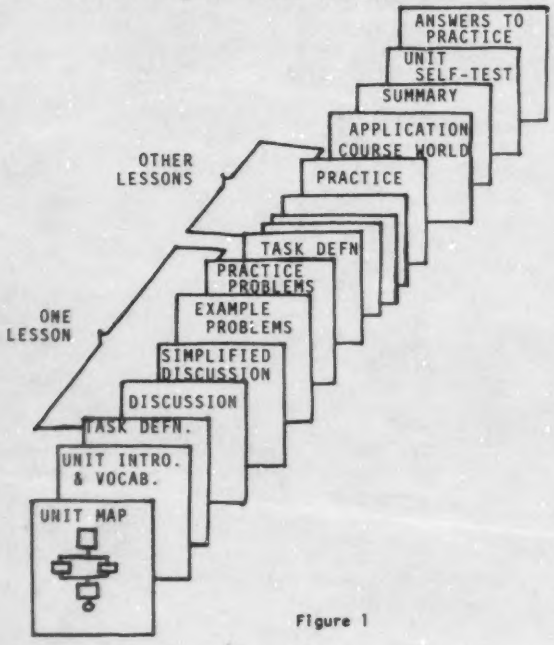


Figure 1

MANAGEMENT SYSTEM

Lectures

At the beginning of the course the student is told that the course is set up for self-study and self-pacing (within certain time guidelines); but for those students desiring the security of a regular lecture format, a regular lecture session would be available. To accomodate the approximately 600 students enrolled each semester, four lecture sections are provided and a detailed daily lecture schedule is laid out for the student. These sections are taught by full-time faculty who have special interests in teaching.

Testing Center

Brigham Young University has a facility that is not standard on all university campuses. Some years ago a testing center was established independent of any of the various academic departments and colleges on campus. The center not only administers the typical national competency and aptitude tests for students but it also serves as a testing center for any department on campus which desires to make use of its services. It has seating to accomodate around 325 students at a time. It is open 8-12 hours a day and on Saturdays. This provides students the flexibility to schedule their own time for taking tests. In-class time isn't used for testing and can be used for instruction. Monitors are trained to oversee the testing operation and are always present.

Their operation has been computerized so that for an objective-type test, the student can receive immediate feedback (within seconds) on the exam indicating which questions were missed, what answers they chose on the missed questions, and even the correct answer if desired. We use this facility in the following manner.

For each test topic we have created 4 parallel exams. We have divided the course so that we have 8 test topics plus a comprehensive final. A student must challenge each test topic but may rechallange each test one additional time at their own expense to improve their grade. The computer keeps track of which exam is challenged and ensures that a different parallel exam is taken on the rechallenge.

In addition to the immediate feedback the student receives, our department receives a weekly alphabetical listing of all students and their scores on all exams up to that point in time. Thus, a regular monitoring is possible for all students, and those having problems can be quickly identified.

To maintain some structure in a system that easily lends to procrastinating, a series of deadlines is set by which each exam must be taken in order for the score to count in the students total. Other than this restriction, a student can go at
his own pace, being able to work around other, less flexible courses in his schedule.

Test Feedback

One of the most useful teaching devices is a good review of the test that has been taken and failed by the student. In light of this observation we provide a test feedback center. Here a student may leisurely examine a copy of the exam he took, and consult with a graduate teaching assistant concerning the how's and why's of his incorrect answers. This area is also closely monitored to protect the confidentiality of the exams.

Homework Centers

We provide a study room manned by teaching assistants practically eight hours per day, five days per week plus several evening hours. Here students can study and receive help on questions

they encounter in their regular reading and problem solving. Test questions cannot be discussed here, but help is given on practice problems found in the text.

Media Helps

Our university library houses a collection of professionally marketed audio cassettes and filmstrips on all topics with several series in statistics. Equipment is available so students may go and view these materials on an individual basis as needs arise.

Instructor

Each instructor reserves office hour time to consult one-on-one with students who desire such consultations.

Extra Credit Options

Students are encouraged to read pertinent articles from the Statistics: A Guide to Unkown series. Encouragement comes by providing extra credit points for such involvement. Also students are encouraged to do a project paper involving data collection, summarization, and analysis. This is written up, typed and submitted for grading. We also provide the opportunity on an extra credit basis for students to work on a time-sharing computer system using Minitab. It is felt that all these options provide extra insight to justify the extra credit given for them.

All of these components work together as diagrammed in Figure 2 to move the student through a successful experience in what is usually a first and only encounter with a formal statistics course. Subjective feelings from unsolicited student comment has been positive and reassuring concerning the textbook and the operation of the course. More objective feedback is reported in the next section.

EVALUATING THE COURSE

Because of the problems of constructing a good experimental design in an ongoing program serving 600 or so students each semester, we do not have the kind of data that would allow us to compare our current operation with our previous operation of the course.

On an on-going basis, the only available information to make comparisons and evaluations is grade distributions. To remedy this situation to some extent, the last few years a series of questionnaires have been administered in the course in order to get descriptive information and opinions about the course.

Summarizing some of this information, a comparison of the grade distributions between Winter Semester, 1973 and Winter Semester 1979 shows a shift in grades towards the A grade although there is hardly any difference in the two years in the total percent of A, B and C's given. This grade shift was anticipated because of allowing tests to be rechallenged. The percent of students receiving D, E, and UW (unofficial withdrawl) grades is not appreciably different for the two years compared. (One concern we had was that the percent of UW's would increase due to the flexibility of the program and the tendency of students to get behind. The data doesn't support this concern with the percent of UW's being the same in the two year comparison.)

The questionnaire data show a minor tendency for more freshmen and sophomores to be registering for the course, though a statistical test declared the change to be non-significiant. About 52-54% of the students declared themselves as attending the lectures frequently or almost always. We infer that more than 40% of the students completing the course are doing it mainly on their own. In addition, the data showed no association between attendance at the lecture and student GPA. There are still questions here that merit further investigation, however.

When asked to rate the text on how well it helped them prepare for the exams about 70% rated it high to very high, about 20% medium and the remaining 10% rated it low to very low. When asked their agreement with the statement that the text presented the material in a clear and concise manner, 82% agreed, 18% were neutral, and 5% disagreed.

For the statement, "the graded tests effectively measured my knowledge," 48% agreed, 27% were neutral and 25% disagreed. Some additional improvement could be made here, though personal experience suggests that students are critical of most tests.

For the statement, "I feel I am receiving a sufficient amount of individual help in the course," 42% agreed, 13% disagreed, and 45% were neutral or unable to judge. This would indicate a lot of students are either not taking advantage of the helps offered or our help is deficient.

Students were asked at the end of the semester to classify their interest in statistics before taking the course and then to indicate their present interest. Fifty-one percent classified their present interest in statistics as "interested" or "highly interested." When looking at the two questions together a shifting of interest before and after the course was observed. Overall, 38% showed an increase in interest while 21% showed a decline in interest.

When asked what rating they would give the course to other students, 38% gave it a "high" to "very high" rating, 41% a "medium" rating and 20% a "low" to "very low" rating. However, when the "interest in statistics" variable is brought into the picture, it was found that of the 51% of the student who expressed interest in statistics, 66% of them gave the course a "high" to "very high" rating.

Several other "affective" questions were asked of students taking their last exam. On a seven-point useless-to-useful scale, 83%

regarded statistics as useful in applied research, and 73% felt it to be useful in their own field. In addition, 69% expressed confidence in their ability to understand an introductory statistics text.

Admittedly, the above percentages are quite meaningless without something with which to compare them. However, in most cases they lean toward the positive side of 50% which we interpret as good signs in the absence of comparative data. We also now have a baseline to which we can compare as other innovations and changes are implemented in our programs.

Relative to the seven problem areas identified before our program was implemented, and listed earlier in this paper, the following progress can be noted:

1. Current student reaction seems to indicate that students see a relevance of statistics to their major field, 73% agreeing with that conclusion.
2. We feel we are seeing more interest expressed in the course, but we are still looking at ways to improve our course in regard to this.
3. There aren't any strong signs that a major shift is taking place that would indicate more sophomores are taking this "sophomore level" class. There are some minor signs and time will tell.
4. We have at most 4 to 5 different instructors teaching our 600+ students each semester and in some cases only 2 or 3, compared to 10-15 under the old program.
5. All instructors are full-time faculty or graduate students carefully selected because of outstanding teaching skills.
6. With the testing center available, we have adequate lab space.
7. The lab manual has been done away with and the lecture, text and examinations have become highly correlated.

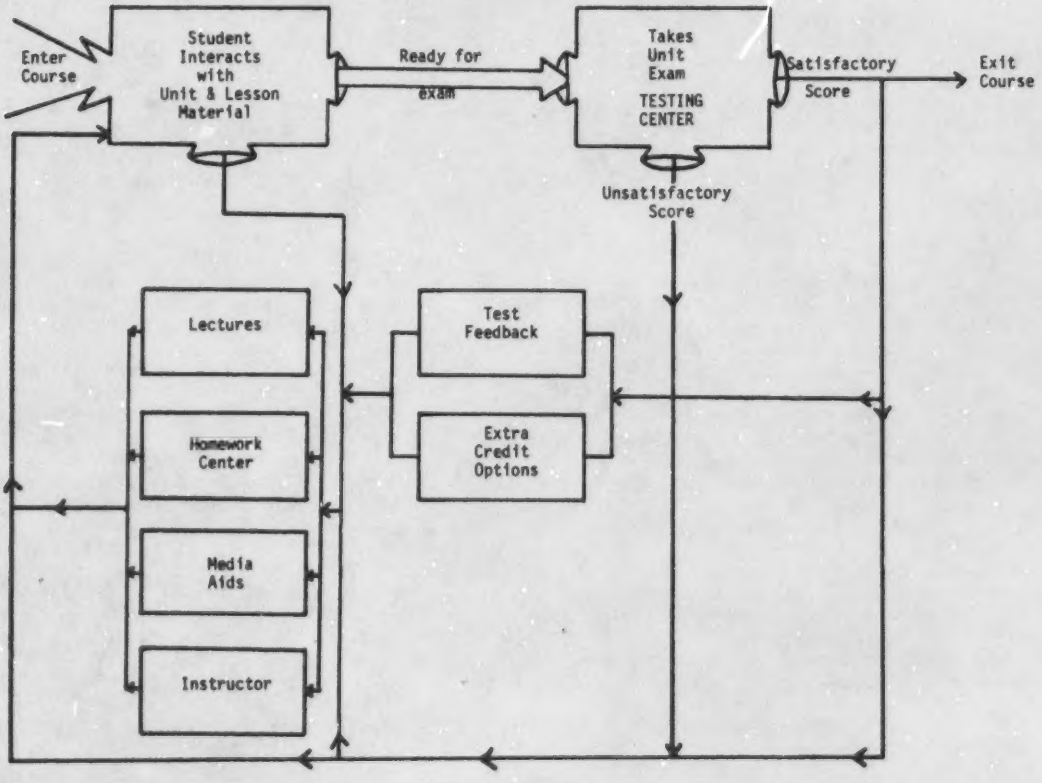
We have areas in which to improve and we continue to work in these areas, but we feel we have made some significant progress by revising our introductory course as we have. Suggestions and criticisms from all sources are welcome.

REFERENCES

Christensen, H.B. and Statistics Instructional Development Team, 1977. Statistics Step By Step, Houghton Mifflin Company, Boston, Mass.

Green, E.E. and Christensen, H.B., 1977. Introductory Statistics: A Simplified Approach, Aspects of Educational Technology, Vol XI, Kogan Page, Publisher.

Figure 2
INSTRUCTION AND MANAGEMENT SYSTEM



THE ELEMENTARY STATISTICS COURSE -- FACT OR FICTION

J. Leroy Folks, Oklahoma State University

1. Introduction

The London Transport Guide to Greater London describes a Roman bath just off the Strand. In fact, several Roman baths have been found in London. Some are described thus: Definitely Roman but is it a bath? The others: Definitely a bath but is it Roman?

At Oklahoma State University we are attempting to teach with a limited number of faculty members an incredible number of students enrolled in pre-calculus service courses in statistics. These courses tend to be one of two types: "definitely elementary but is it statistics?" or "definitely statistics but is it elementary?"

During the fall and spring semesters we enroll about 1300 students each semester in pre-calculus service courses. About 850 of these students are registered in sophomore-level courses for non-majors with little mathematical background. Most of the comments which follow are prompted by my attempt to evaluate our efforts with these students but perhaps my comments will be of interest and value to those of you at other universities.

2. Questions Suggested by Fact or Fiction Part of the Title

We have just completed revising a description of our courses. It therefore seemed appropriate to me to ask whether these descriptions are factual or largely fictitious. I have therefore posed a number of questions:

a. Can we teach statistics to students not yet involved in research?

My experience in attempting to do so prompts me to say that it is very difficult and observation of my colleagues' attempts leads me to the same conclusion. P.A.P. Moran [1] in his paper entitled "What Should a Professor of Statistics Do?" says "The real problem here is not statistical method, but statistical reasoning, which is difficult to teach. This is not a matter of formal mathematics, or of learning how to calculate statistical coefficients. What is necessary is to teach what the results of such numerical calculations imply or suggest."

Walt Federer [2] in his paper entitled "Some Remarks on Statistical Education" begins by defining statistics to be concerned with the characterization, development, and application of techniques for

i. the statistical design of an investigation whether it be an experiment, a survey, an observational study, or a model building study;
ii. the summarization of the facts from the investigation; and
iii. the inferences that can be drawn from the facts of the investigation about the parameters in the population.

Without a prior exposure to research of some kind, a student will find the concept of the statistical design an abstract idea indeed!

b. Does it matter whether we use real or imaginary data for exercises and examples?

I think that it matters but not as much as we might imagine. Let me explain. I have found students to be generally passive with respect to the realism of exercises and examples. What seems to be more important is that they encounter examples in their precise field of study. A wildlife management student studying species of plant life may not be able to see the worth of an example using different species of animal life. A home economics student may have difficulty in relating an experiment in agronomy using different levels of fertilizer to a baking experiment using different oven temperatures.

Data which might be real are helpful but, more importantly, the data should be real to the student. I have found data actually generated by the students themselves in classroom experiments to be the most helpful.

c. Can we teach statistics to large sections of 250 students?

During the last year we began to use very large lecture classes (250 students) while giving true-false/multiple choice questions on computer graded examinations. This has to be regarded as experimental as we are not yet sure that this makes sense. In an operational sort of way, the answer to my question has to be yes because students did attend, were given examinations, and were given final grades in the course. However, I can not help but wonder if it is somewhat like teaching piano to 250 students. One could give lectures, administer true-false examinations and give final grades -- all without sight or sound of a piano. Some students would receive A's, some B's, etc. But none of them would have been taught to play the piano.

Of course, one can argue that if they are able to read the computer sheet they receive after each examination, that they have learned some statistics. Nevertheless I feel compelled to agree with Moran that teaching statistical reasoning is difficult. It is even more difficult in large lecture classes.

d. Can we use statistics graduate students and non-statistics faculty members to teach elementary statistics courses?

My answer to this question is a qualified yes. Graduate students are themselves closer to the problems of studying and learning and so, in one way, are better qualified than the older faculty members to teach statistics. On the other hand, my experience has been that graduate assistants tend to make mathematics courses out of elementary statistics courses. This is understandable since they are not well grounded in the history and philosophy of statistics and have not yet mastered statistical reasoning.

We now have considerable experience in having statistics courses taught by non-statistics faculty. On the whole they do a good job but there are some problems.

(1) They may present a very narrow view of statistics.
(2) They may not keep contact with the statistics faculty.

(3) Their departments begin to believe that they have control of the course--not the statistics department.

e. Are there really people interested in teaching the elementary course?

Yes, for a few times at least. Yet there are few who wish to devote their careers to such courses and the many innovations may indicate a dislike for teaching the course. At least we don't see so many innovations with other courses.

3. The Fact-Fiction Aspects of a Typical Elementary Book

Most of us specify textbooks, the students buy the books, and we assign reading for the students. Yet nearly all of these books, including the one I am trying to write, have defects. Some of the defects constitute terminal illnesses; in other cases the ailments are minor.

Important material is often omitted and much of the material included is false at the worst, controversial at best.

a. People and Places

(1) The Neyman-Pearson Story

A few years ago at a social gathering at a statistics meeting, several of us were discussing the teaching of business statistics courses and the choice of textbooks. I indicated that the viewpoint of one particular book was more Fisherian than it was of the Neyman-Pearson school. One of the people present who had been teaching statistics indicated that he was not familiar with Neyman and Pearson, apparently believing them to be authors of some current textbook in business statistics. To me this was somewhat like a teacher of the Old Testament saying he had never heard of Moses.

(2) The Rothamsted-Ames Story

Ignorance of the important people and places in the early days of statistics is deep. In the manuscript which I have prepared for an elementary course in statistics I have included a short chapter about Fisher's days at Rothamsted and Snedecor's work at Iowa State University. I chose to call this chapter The Legacy of Rothamsted and Ames. Recently I have been surprised by the number of people educated at excellent departments who are unfamiliar with Rothamsted and Ames.

(3) The Gosset Story

Sometimes historical elements are included in a book which are almost true. In several books I have looked at recently I have seen statements that Gosset published the distribution of $(\bar{x} - \mu)/s/\sqrt{n}$ and this has been called Student's t ever since. Strictly speaking, Gosset did not publish the distribution of t in 1908 but that of $\sqrt{n-1}\,t$ and t was not used until much later. Churchill Eisenhart [3] gives a good presentation of the transition from Student's z to Student's t.

b. The Statistical Design

I am in agreement with Federer's definition of statistics and lament the fact that so many books convey wrong ideas concerning the statistical design. To cite a few examples:

(1) The experimental unit is something on which a measurement is taken. This is, of course, not true and would fail to make a distinction between an experimental unit and a sampling unit.

(2) A completely randomized design is sampling from independent populations. This definition leaves no room for treatments, experimental units, or random assignment of treatments to the experimental units.

(3) Randomization of treatments guarantees independence of errors. As is well known, randomization does not guarantee independence or even zero correlation. It does induce equal correlation among errors.

c. Puzzles

Sometimes we have ways of saying things unique to statisticians which are perpetuated in the belief that some purpose is being served. For example, some books still insist that the following two statements are different:

(1) the interval contains μ

(2) μ is in the interval.

It is usually argued that one statement implies that the interval is random; the other statement implies that μ is random. I do not believe that this argument can be supported by the usual rules of English grammar. Neither do the students, apparently, for they always have trouble in seeing the difference.

A teaching which I personally find very objectionable is that if $n > 30$ we do one thing, if $n < 30$ something else.

4. My Own Attempt

I have been a little caustic in my criticism of elementary statistics textbooks. In case anyone here has been offended, I hope that they will have the same opportunity in the future to criticize me. I have been working on an elementary statistics manuscript which I hope will be published. In this manuscript I have tried to emphasize the historical thread in statistics in the U.S. and Britain, particularly from 1850-1950.

a. The Setting

In the late 1800's Britain sat at the peak of empire, the wonder of the civilized world with an air of optimism and a belief in the possible achievements of science. The United States had nearly expanded to its continental boarders, the Civil War was over and an unprecedented age of technology was about to begin.

b. The Participants

Into this setting there came a relatively small band of people who virtually accomplished miracles in the development of what we now call statistics. To name a few of these, K. Pearson was born in 1857, gave his lectures on probability and statistics during the years 1891-94, was the co-founder of Biometrika in 1900, established the Department of Applied Statistics in 1911 at the age of 55, founded Eugenics in 1925 at the age of 68, and retired in 1933 at the age of 76!

George W. Snedecor, was born in 1882, was appointed assistant professor of mathematics at Iowa State in 1913. In 1925 he began a series of Saturday afternoon seminars for faculty members. In 1931 he invited R.A. Fisher to Ames and

UU 48[illegible]

Gertrude Cox received her M.S. degree. In 1937 he published Statistical Methods.

R. A. Fisher, born in 1890, graduated from Cambridge in 1912, the same year that he published his first paper. He was employed at Rothamsted Experiment Station from 1919 to 1933 and was Head, Department of Eugenics, University College London 1933 to 1947.

John Bennet Lawes, born at Rothamsted, patented superphosphate and built a factory at Deptford Creek near the old Royal Dockyard. He left a sizable endowment for the Rothamsted Experiment Station.

c. The Great Ideas

I have tried to emphasize some of the great intellectual achievements:

(1) Mendel's work.
(2) Use of mathematical equations and curves to describe data.
(3) Student's small sample statistics.
(4) Fisher's work at Rothamsted.
(5) Neyman-Pearson theory of testing.

d. The Great Controversies

(1) The great Bayesian controversy.
(2) The great Fisher-Pearson controversy.
(3) Inference vs. decision.

5. Conclusions

a. Our elementary statistics courses do exist in fact but it is a fiction to believe that it is easy for them to be effective.

b. Our objectives must be modest.

c. The courses don't have to survey all of the topics in statistics.

References

1. Moran, P.A.P. 1975. What should a professor of statistics do? Australian Journal of Statistics, 17, 121-133.

2. Federer, Walter T. 1978. Some remarks on statistical education. The American Statistician, 32, 117-121.

3. Eisenhart, Churchill. 1979. On the transition from "Student's" z to "Student's" t. The American Statistician, 33, 6-10.

TEACHING STATISTICS TO MEDICAL SPECIALISTS - CARDIOLOGISTS

James H. Ware, National Heart, Lung, and Blood Institute

When preparing to teach statistics to a group of medical specialists, consider the work environment for members of your audience. Typically, the medical specialist has a heavy schedule of responsibilities and many commitments competing for his time. Lectures, for example, are likely to be scheduled late in the day to avoid conflicts. Unlike the graduate student or undergraduate medical student, course work is not a primary responsibility. Typically, your students are there because they believe that they need to know more about statistics. Thus, there is an increased need to present ideas as directly as possible, i.e., without a drawn out development, and to make the material relevant. At the same time, there is an opportunity to teach to an especially responsive and capable audience.

Another issue to consider is heterogeneity of the group with respect to statistical background. Although few in the class will have significant recent background in statistics, the level of sophistication about the quantitative ideas and recall of statistical ideas will vary considerably. Thus, it is useful to present material that can be appreciated at a variety of levels.

Certainly there is no single correct way to teach statistics to medical specialists. However, I would like to propose six ideas which I have found helpful in teaching statistics to medical specialists and hope that others will find some of these ideas useful as well.

1. Use the Case Study Method

The easiest way to get off to a good start with a group of medical specialists is to present an intriguing problem, from your own consulting experience if possible, and otherwise from the medical literature. For example, I was recently analyzing data from a series of patients who had undergone operation to remove a portion of the enlarged wall between the right and left ventricles of the heart. The condition is known as asymmetric septal hypertrophy (ASH) and becomes clinically significant when the enlarged wall begins to obstruct blood flow in the heart. The pressure gradient across the obstructed area was measured both before and after operation.

We compared the impact of operation on various subgroups of patients by the change in gradient after operation. In other words, the analysis was based on the difference between post-intervention and pre-intervention values. This design, as you know, is often called using the patient as his own control, and is very common in the cardiology laboratory.

As it happened, the impact of operation was so great that the pre and post values were very weakly correlated. In fact, a simple comparison of the post treatment values was a more sensitive analysis of intergroup differences! Although the ideas arising in this example are somewhat advanced and might arise late in a series of lectures, I plan to use this example the next time I discuss designs with baseline measurements and their value relative to simple randomized experiments. Given the time, I would use the example to motivate the analysis of covariance.

Good examples will sustain the interest of your audience and minimize problems arising from varying statistical background. They tend to identify problems of common interest and stimulate discussion. And, by using the data (in a highly simplified form), fundamental ideas like calculating summary statistics can be reviewed and more difficult computations can be illustrated.

2. Consider the Difference Between Training and Educating

I have found this distinction helpful in defining my goals in teaching courses to medical specialists. We define the two verbs for this discussion as follows:

To Train: To impart a set of skills by instruction or practice.

To educate: To impart general knowledge and enhance the powers of reasoning or judgement.

As statisticians, we tend to feel that the conceptual basis of statistics is an essential part of the subject. Nevertheless, I have had great success when teaching the specific skills needed to carry out an analysis. Thus, when a good example has shown the utility of the correlation coefficient as a summary statistic, the student can learn how to compute a sample correlation coefficient and assess its significance level in five minutes. Although I certainly do not suggest avoiding fundamental concepts like sampling distributions, I do suggest that many courses and text books devote a considerable time to laying down a foundation without ensuring that the material is motivated. What I am emphasizing here is the need to move from the specific to the general in a way that seems relevant to your audience.

3. Find the Right Level of Difficulty

This can sometimes be a problem in this setting. Some members of the audience may be working with statistical concepts on a daily basis in their work while others are beginners. As I have mentioned, interesting examples help solve this problem. In the context of examples, basic ideas like standard derivations, error rates and sampling distributions can be thoroughly reviewed, while a prolonged introduction of these ideas in isolation would be less likely to hold the interest of the audience.

On the other hand, it is easy for the mathematically trained instructor to move too fast or formulate problems too abstractly. When the problem is well motivated, the medical specialist will want to see the solution of the problem in careful detail. Here again, I find repetition of the analysis of the sample problem a useful tool.

4. Interact with the Group

The medical specialist usually has a clear idea of what he wants from the course. By inviting feedback during and especially after

GUC 50

class, you can find out if your examples are interesting, your pace appropriate and if the students are finding the material worthwhile. Even the best course can be improved in this way, and I've sometimes made important changes in a set of lectures on the basis of feedback from my students.

5. Don't Try to Do Too Much

Often, you may be asked to teach several classes on statistics in the context of an interdisciplinary course. In this context, I try to treat a few ideas in detail rather than give an overview of basic statistics.

6. Consider Doing Some Calculations with the Group

Because the working medical specialist has many demands on his time, homework assignments may not be a useful teaching tool. In this situation, an example can be started toward the end of the class, possibly with some calculations left to the participants. If the framework is laid out and the example partially completed, students are more likely to finish it later.

7. Some General Remarks

Medical specialists are often fun to teach. They are bright, well-trained and have good instincts about experimentation. Also, they are likely to be highly motivated when participation is voluntary.

They are easier to teach when their medical and statistical background is homogeneous, an ideal rarely achieved. Some of the ideas I have discussed help to solve problems arising from diversity.

Practical experience, especially in a specialty area well represented in the audience, is a boon to the instructor. Courses in medical concepts for graduate students in statistics may be of value, but they are not a substitute for extensive collaborative experience.

Finally, medical specialists are very selective in how they use their time. Even a first rate job may fail to hold the interest of some of your students. Nevertheless, the rewards from this kind of teaching can be great. I have always learned a lot about statistics, medicine and the collaborative process from teaching to medical specialists, and I look forward to doing it again.

DISCUSSION

Dennis B. Gillings, University of North Carolina at Chapel Hill

Health services problems are multidimensional and a good deal of current research is data-based and so requires statistical input. Experiences plus the skills of combinations of economists, sociologists, administrators, physicians, nurses, dentists, lawyers, epidemiologists, statisticians, operations researchers, and other health providers are typically required. The methodological problems in health services research are usually considerable, and so there is great need for training to include advanced statistical design and analysis skills. In my experience, however, statisticians are not generally turned on by health services (say, in contrast to clinical trials) because, although the methodological problems are great, the problems of definition, measurement, data collection, practicality, timeliness, and political expediency are even greater, and they tend to prevent one embarking on a neatly designed study. Instead, a patchwork of avenues are often necessary that can make the analysis rather messy.

I have myself estimated that between 400 and 500 health services oriented statisticians are needed in the United States during the next five years. This estimate is derived from the needs implied by legislation in the 1970's concerning quality control of health services, the Cooperative Health Statistics System and Health Planning and Resources Development, and by the implications of imminent legislation concerning national health insurance. Note that there are about 200 Health Service Areas and another 200 PSRO areas for quality control, and so these estimates may be conservative.

Two types of health services specialists can be identified for training purposes. First, the health services researcher who needs some basic training in statistics. Second, the statistician whose field of application is health services research. During the past eight years in the Department of Biostatistics at UNC, Chapel Hill, I have dealt with both types and feel there are several important features that deserve mention.

A major difficulty with training health services researchers of either type is the risk of developing a jack-of-all-trades and master of none, versus a disciplinary specialist who is of no practical use. We tackle the problem at Chapel Hill by training health services researchers within the separate discipline (e.g., epidemiology, health administration and biostatistics) and exposing the students to more general considerations through core courses and special seminars given by the UNC Health Services Research Center.

Another general issue is the experience of the student. Trainees need at least one or two years (preferably three or more) working on health services related problems or in health agencies in order to get much out of a health services training program; otherwise, they tend to be involved only insofar as they learn about an individual discipline. Major health services problems today concern costs, manpower distribution, a highly fragmented system, excessive specialization, excessive technology, and access. The problems are vast and complex, requiring considerable knowledge before they can even begin to be tackled. And so, the instructor of statistics faces a major problem, because he or she needs to know about selected health services problems in some depth. There are few statisticians who have been trained to be health services researchers. Instead, knowledge about health services is likely to come from past experience. It is important that this experience gets translated into examples and case studies that are presented to a class of students. In fact, the examples are the crucial feature of these sorts of statistics courses (no different to any other medical specialty!). However, the experience of the instructor need not be of any particular type. It could be related to planning, evaluation, or health policy. Topics concerning health manpower, utilization, health needs, quality of care, hospital service areas, health maintenance organizations, health insurance, health technology, or a myriad of other items are relevant.

Perhaps the biggest headache facing the teacher of statistics to the health services student who is not a statistical specialist is that students need to be adept with an array of multivariable and multivariate techniques in order to solve the problems they are facing, but it is not easy to teach multiple regression, path analysis, factor analysis, cluster analysis, and MANOVA to students who have not adequately grasped (or been exposed to) the basic statistical courses. One really needs to be able to train a general data analyst who understands the weaknesses and strengths of available statistical methodology and who is thoroughly trained in a computer oriented approach. It is not easy, perhaps impossible, to teach what I call the fundamentals of data structure without training a statistician per se. By fundamentals of data structure, I mean recognizing the intricacies and implications at the data analysis stage of the

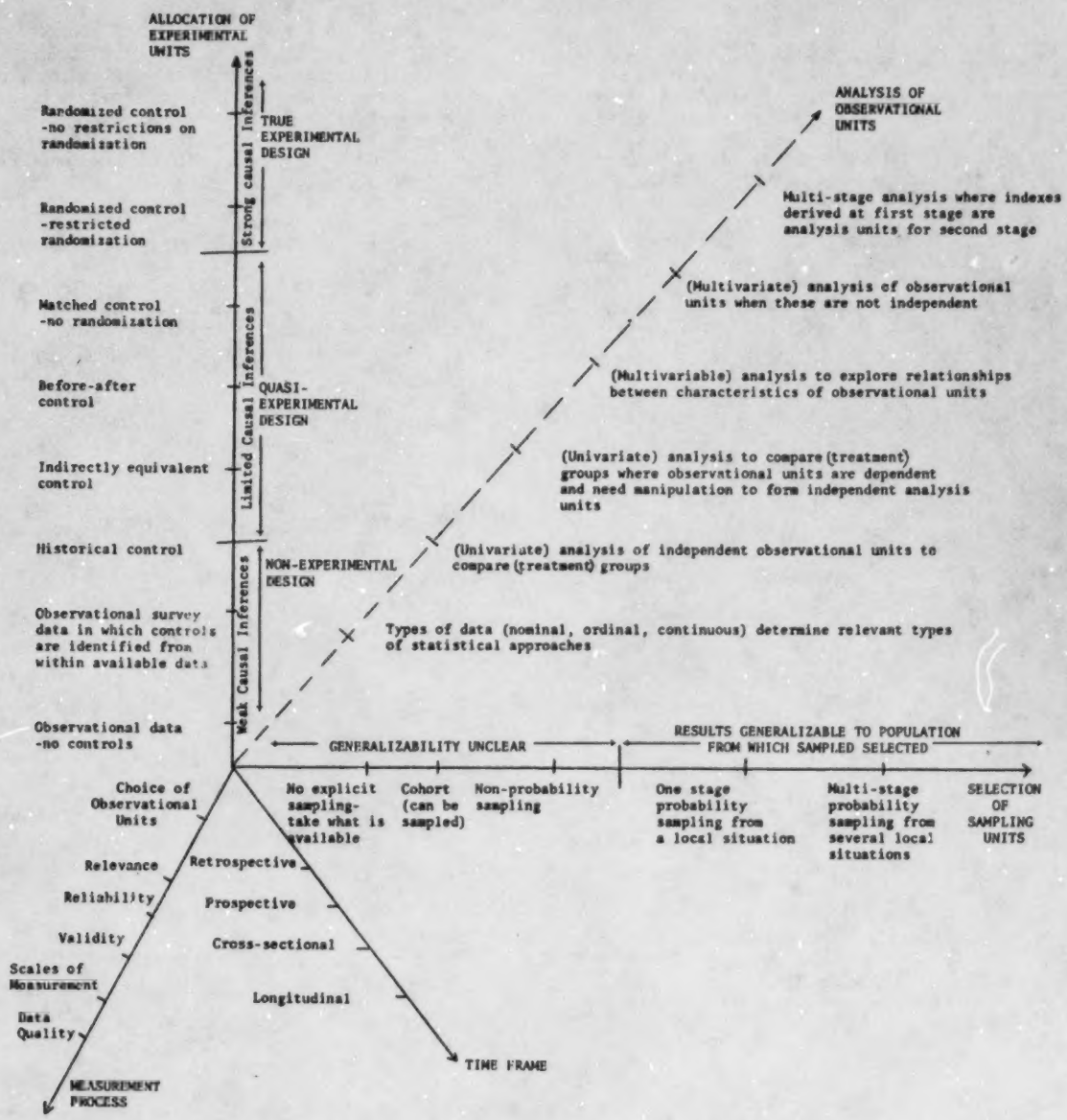
1) measurement process,
2) sampling process,
3) study (or experimental) design, and
4) time frame.

Easy familiarity with sampling units, experimental units, observation units, analysis units is critical. But the full breadth of these notions is not typically included in courses (probably because few people know how to do it), and the armory that applied statisticians use to disentangle any given situation is built on a foundation of specialized knowledge. The chart in Figure 1 summarizes the notions I am alluding to. The analysis axis makes a fifth dimension and in a typical health services study, we may be anywhere in the implied five dimensional space. Studies are likely to have several components, often in different parts of the space. Connecting what sometimes appears to be disparate features of studies is difficult for the expert data analyst and so, understandably, a prodigious task for the health services researcher.

Figure 2 attempts to identify the major methodological skills that are required by the health services researcher. Only a methodological or applied statistics specialist would get a

UU: 52

Figure 1 The underlying design and measurement structure that generates a data set for analysis


ALLOCATION OF EXPERIMENTAL UNITS
Randomized control -no restrictions on randomization
Randomized control -restricted randomization
Matched control -no randomization
Before-after control
Indirectly equivalent control
Historical control
Observational survey data in which controls are identified from within available data
Observational data -no controls
Strong causal inferences
TRUE EXPERIMENTAL DESIGN
Limited Causal Inferences
QUASI-EXPERIMENTAL DESIGN
Weak Causal Inferences
NON-EXPERIMENTAL DESIGN
ANALYSIS OF OBSERVATIONAL UNITS
Multi-stage analysis where indexes derived at first stage are analysis units for second stage
(Multivariate) analysis of observational units when these are not independent
(Multivariable) analysis to explore relationships between characteristics of observational units
(Univariate) analysis to compare (treatment) groups where observational units are dependent and need manipulation to form independent analysis units
(Univariate) analysis of independent observational units to compare (treatment) groups
Types of data (nominal, ordinal, continuous) determine relevant types of statistical approaches
GENERALIZABILITY UNCLEAR
RESULTS GENERALIZABLE TO POPULATION FROM WHICH SAMPLED SELECTED
Choice of Observational Units
No explicit sampling- take what is available
Cohort (can be sampled)
Non-probability sampling
One stage probability sampling from a local situation
Multi-stage probability sampling from several local situations
SELECTION OF SAMPLING UNITS
Relevance
Reliability
Validity
Scales of Measurement
Data Quality
MEASUREMENT PROCESS
Retrospective
Prospective
Cross-sectional
Longitudinal
TIME FRAME

NOTE: Any situation which occurs in practice may include a combination of categories along any one dimension as well as a combination of up to all five dimensions.

reasonably deep knowledge of most of these items, but the general health services researcher does need exposure to many of the topics because they will occur frequently.

The greatest impact can be obtained by carefully choosing the way in which statistics is taught. Heavy emphasis needs to be placed on case studies, field training, consulting problems, technical assistance to health agencies, and report writing. Further data management and data sources are important topics that require emphasis but these topics are not usually undertaken with relish by statisticians. Heavy emphasis needs to be placed on statistical packages as the health services researcher, planner or evaluator will typically have moderate to large data sets, possibly with a large number of variables. There is little need to teach data analysis using a calculator as it is doubtful that such procedures will be employed in real studies.

Survey methodology also needs heavy emphasis, as well as quasi-experimental designs covered in the book by Campbell and Stanley. Statisticians often give short shrift to the non-sampling aspects of survey methodology, but these are terribly important to health services specialists. Likewise the analysis of quasi-experimental designs can be challenging. However, it is of little use to expose health services students to rather esoteric experimental designs which can be used only in situations in which the investigator has a great deal of control (e.g., agricultural, engineering, industrial).

Really, statistics needs to be taught in an integrated way so that disciplines and topics such as epidemiology, research design, and evaluation are woven into the substance of the course. This can be very difficult for one instructor. Unfortunately, the problem is rarely solved by scheduling a procession of instructors for different subjects as then it is terribly difficult to maintain continuity and integrate the material.

In summary, the following brief recommendations may help considerably in the statistical training of health services specialists.

1) Use many varied examples related to health services issues for presentation of statistical methods.

2) Consider the use of a case studies course.

3) Emphasize heavily student involvement in real projects under faculty supervision.

4) Require at least one or two years experience before graduation, preferably before admission.

Figure 2

TECHNIQUES AND METHODS - A SUMMARY

DESIGN			INFORMATION SYSTEM		ANALYSIS	
SURVEYS/ OBSERVATIONAL STUDIES	EXPERIMENTS/ CONTROLLED STUDIES	MODELS	DATA COLLECTION	DATA MANAGEMENT	DATA ANALYSIS	DECISION ANALYSIS
1. Sampling I. Non-probability samples i. Quota sample ii. Convenience sample iii. Judgment sample-Delphi Technique II. Probability samples i. Simple random sample ii. Systematic sample iii. Stratified random sample iv. Cluster sample v. Multi-stage samples 2. Questionnaire Design Design of Data Collection, Instruments 3. Interviewing Techniques 4. Survey Management	1. Pre-experimental designs i. One-shot case study ii. Static group comparison 2. True experimental designs i. Pre-test, post-test control group design ii. Solomon 4 group design 3. Quasi-experimental design--e.g. Time Series Design 4. Prospective and Retrospective Studies, Historical Prospective Studies 5. Natural Experiments 6. Clinical Trials 7. Other more sophisticated true experimental designs i. Latin Squares ii. Randomized Blocks iii. Fractional Designs, etc.	1. Mathematical functions and equations 2. Distributions 3. Mathematical models for probability sampling 4. Mathematical models for other observational studies--e.g. regression, path analysis 5. Mathematical models for experiments--e.g. ANOVA 6. PLNT, GERT 7. Goal Fulfillment Model, MBO, PPBS 8. General systems models, feedback loops 9. Programming models (linear, quadratic dynamic, integer) 10. Queuing models 11. Markovian models 12. Life tables 13. Other mathematical models for data analysis	1. Ask questions i. Interviews--personal or telephone ii. Mail--cheaper but lower response rate 2. Read i. Content analysis of documents (item not recorded does not mean it did not take place) ii. Refer to other reported data--e.g. data sources NCHS, Census Bureau, etc. 3. Observe--e.g. time motion studies--need highly trained observers 4. Data from secondary source--e.g. NCHS, Census, other private and governmental agencies 5. Design of data collection instruments	1. Hardware 2. Software--Programming language 3. Method of data entry 4. Method of data storage 5. Method of data retrieval 6. Method of updating 7. Record Linkage 8. Data Editing 9. Quality Control of data 10. Data analysis requirements	1. Description i. tables, rates, adjusted rates ii. graphs, charts iii. central tendency iv. dispersion v. distributions vi. correlation 2. Estimation i. Method of moments ii. Least Squares iii. M.L. iv. Simulation 3. Test Hypotheses Z,t,χ^2,F tests 4. Special Techniques for Analysis of Models--for example i. Estimation techniques for sampling ii. ANOVA, MANOVA--general linear model, includes regression 5. Special techniques for ordinal or nominal data i. Non parametric techniques ii. Categorical Data Methods 6. Multivariate methods--in particular, Factor analysis, Discriminant analysis, Commercial Correlations, Cluster Analysis, Path Analysis	1. Benefit-Cost analysis, Cost-Effectiveness analysis 2. Decision Trees, Utility, Probability, Sensitivity Analysis 3. Experiment with models 4. Forecasting techniques
PRINCIPLES AND OTHER CONSIDERATIONS Probability Random Selection Validity, bias Reliability Non-response	Replication, Matching Randomization Control, Blind, Internal Validity External Validity Ethical problems	Fit of model Is model tractable? Hypotheses to be tested? Mathematical models often have a bad image	Secondary data should be used where possible	With large amounts of data, general management problems become important	Data analysis packages SPSS, Biomed, etc. Types of data--nominal, ordinal, interval. Selection of approach from vast amount of techniques available	Decisions are not usually made using these techniques. Estimation of probabilities & utilities can be difficult in practice
GENERAL DESIGN CONSIDERATIONS STATE PROBLEM--SET OBJECTIVES--DESIGN STUDY Develop budget--allocate resources to each phase of study			GENERAL INFORMATION SYSTEM CONSIDERATIONS Quality and timeliness of data Confidentiality Accessibility Uniformity and comparability of data		GENERAL ANALYSIS CONSIDERATIONS There is often a lack of communication between the analyst and the administrator due to the complexity of methods involved.	

A COURSE IN ELEMENTARY INFERENTIAL STATISTICS THAT STUDENTS UNDERSTAND

David Lile Brown, University of Connecticut

This paper describes a course called Nursing 260, Statistical Analyses for Research in the Health Professions, which is taught by the author at the University of Connecticut. This paper discusses the need for such a course, the way the course was developed, how the course operates, and evidence on how well the course works.

The Need for Such a Course

Many students need a course in statistics to complement their science based curricula. For example, students in nursing and the allied health professions are often required to have a statistics course either as a baccalaureate requirement or as a prerequisite for admission to graduate schools. For these students and many others, the need is for a course which emphasizes applications rather than mathematical derivations and proofs. They need a course that is directed toward the usefulness of statistics to their professions. They need a course they can understand, succeed in, and one from which they will retain knowledge. They need to cover in one semester a full range of basic inferential procedures for categorical, ordered, and interval data.

Unfortunately, there was no such course offered to these students anywhere in the university. Introductory courses in the statistics department cover only one half or another of the material and they emphasize the details of proving mathematically that statistics do work. The complaints that many people make about such basic statistics courses are:

1. The material does not make sense;
2. After a year or so, they have forgotten whatever they learned; or
3. They are confused and still do not know what to do when faced with a problem what requires the application of statistics.

The few students who took one of those courses (few students from the health professional schools would sign up for statistics) indicated that they did not benefit very much from doing so. Students in our graduate program who had had such courses as undergraduates were incapable of deciding what procedures to use for their research projects. Thus there was an obvious deficiency to be overcome.

How the Course Was Developed

Having an understanding of the need for a course in applied statistics, and holding onto the belief that a course could be designed that would meet the need, I suddenly saw how to do it one day. I received in the mail a review copy of a book from Wadsworth Publishing Co. entitled The Practical Statistician: Simplified Handbook of Statistics by Linton and Gallo. The light came on! This was it -- a structured approach designed to help people pick out the important features of any given research design, select and perform the correct statistical procedure, and make accurate conclusions based on the results. The text even covers situations under which the underlying assumptions of the procedures are not met. This gave me the textbook and the approach to inferential statistics that I needed to develop the course.

I set about designing the course by seeing how well the content of the text would fit into a fourteen week semester. I polled students to determine if enough of them were interested for me to be able to give the course a try. I decided that the course could be offered and would fit into the semester very nicely.

Labelled as an independent study course, but offered in a fairly traditional classroom manner, the course was first offered in the spring semester of 1977 to 29 seniors in the School of Nursing. Decisions about the reading assignments, the use of a question and answer lecture format, weekly tests, and immediate feedback were all made with principles of learning theory in mind. Lecture examples and test questions were derived from research relevant to nursing students.

During that first semester, the students responded very well to the material, the text, and the approach I was taking to the subject. They soon declared, however, that they needed homework examples like the ones I prepared for their tests. The text does cover one example in detail for each procedure, but practice examples are lacking. Thus, I had to find more relevant examples, some for homework, and some for tests.

It became obvious that the course would be as successful as I had hoped, so procedures for getting the course officially accepted as a regular offering were begun. During the fall semester of 1977, while a second group of students and three faculty members were taking or auditing the course, a proposal was prepared. It was approved by the School of Nursing Curriculum and Courses Committee, and it was subsequently approved by a vote of the faculty. In the middle of that semester, the course was changed from an independent study to Nursing 260, Statistical Analyses for Research in the Health Professions.

How the Course is Now Taught

I currently offer one section of this course (30-40 students) each fall semester and two sections each spring semester. The text I mentioned above provides the structure for the course. The emphasis is on inferential statistics and experimental designs. Both parametric and nonparametric techniques are covered. Several correlational and specific comparison methods are included and related to the primary inferential methods, and material on using calculators and the history of statistics is added.

The course is taught in one three hour session per week. The first session is spent discussing the need for statistics, the need for this specific course, how calculators can be used to advantage in statistics, and how the course operates. The weekly assignments are augmented with handouts and take-home problems. Each class session, from the second week on, begins with questions from the students about the week's assignment in general. These questions are answered with mini-lectures which are to the point of the questions, but which often

back up to provide evidence of where the answer is coming from. If enough questions are not asked, I ask a few rhetorical questions so that the points I know are most often troublesome do get covered.

Once the material has been covered in general, we start on the homework problems. Solutions are written on the chalkboard and discussed with an emphasis on the thought process behind the various decisions required. Again, students ask questions, I ask rhetorical questions, and several mini-lectures occur along the way.

When all the problems have been solved, and all the questions have been answered, I ask the class if there are any more questions before we have our weekly test. If there are none, the test begins. The first two tests (given weeks 2 and 3) are closed book tests on the first four chapters in the text. These chapters cover terms and definitions which I insist they should know without referring to the book. These are the fundamentals that they must constantly apply throughout the semester.

Tests 3 through 13 (weeks 4 through 14) are open book tests that require students to solve word problems and answer a number of fill-in type questions about the design. Most often a small amount of realistic data is included with the problem, and students are required to compute an actual result and write a brief interpretation of their "findings." Examples are not only fairly realistic, but also relevant to the students in health professional schools since they are taken from health professions research. One intention is to force the students to think about the variables they will encounter in their careers in terms of the research hypotheses that are just begging to be investigated. With the confidence they gain in the course, they learn that they can tackle research problems in a straightforward manner.

Once a student finishes taking the test, he or she hands it in and leaves the room to take a break. While the students are out on break (and usually talking about the test to guess how well they have done), I correct and grade the tests. Out of the 13 tests given each semester, the best 10 scores count toward the final semester average. This allows for missed classes on account of illnesses, snowstorms, or personal reasons. It also allows conscientious students to improve their grades by making up for a poor performance here or there. Students really appreciate this grading policy and work hard because of it. They appreciate knowing just where they stand in the course after every test, and they are forced to keep up with the work. There is no dreaded midterm or final examination.

After the break, I hand back the graded tests, go over the correct answers, and answer students' questions. After all of these questions have been answered, I go over the upcoming assignment for the next week, warning them of various pitfalls and sources of confusion.

Copies of handouts on the assignments and textbook errata are attached as the next few pages of this paper. Also attached is a copy of one week's homework examples, a sample test from the first week of class, and a sample test from later in the semester.

Evaluation of the Course

This course has been very well received and appears to be meeting the needs for which it was designed. A total of 175 students and 13 faculty have taken the course over the past five semesters. Formal teacher evaluations collected by the university have been very positive and seem to indicate high student satisfaction with the course and the way it is offered.

The grading in the course is criterion referenced. It is possible for everyone in the class to get an A or to fail. The tests require that students prove their ability to use statistical procedures correctly. The level of achievement of the students in this course is very high. I consider this an indication of successful learning and an indication that the course works well.

I hope that others will benefit from this work and adapt it to their own purposes at other universities. I will gladly accept suggestions for further improvements in the teaching of statistics.

The appendices originally attached to this paper are too lengthy to include in this proceedings version. Copies of the original paper may be obtained by writing to the author: Dr. David Lile Brown, 36 Mansfield Hollow Road, Ext., Mansfield Center, Connecticut 06250

ON THE USE OF CANONICAL ANALYSIS TO TEST MANOVA HYPOTHESES

Richard K. Burdick, Arizona State University

1. Introduction

The relationship between regression analysis and the analysis of variance (ANOVA) is illustrated in many applied statistical textbooks by the use of indicator variables (see, for example, Neter and Wasserman [1974], Green [1978], and Wesolowsky [1976]). Illustration of this relationship is important because it provides the student with an introduction to the general theory of linear models. Although this relationship can be extended in a similar manner to multivariate statistics, many applied multivariate textbooks either omit discussion of the relationship or mention it without proof or illustration.

This paper briefly reviews how indicator variables can be used to relate regression analysis and ANOVA for the univariate case and then applies the same technique for the multivariate situation. The technique is then illustrated with a real-life data set.

2. The Univariate Relationship

The regression approach for testing ANOVA hypotheses defines a regression model with "dummy" independent variables to represent the qualitative factors of the ANOVA model. For each factor or blocking variable in the ANOVA model, a set of "dummy" or indicator variables is defined to indicate the factor level a particular observation has received. In general, if the j^{th} factor in an ANOVA design containing no interaction effects has k_j levels, $k_j - 1$ indicator variables are defined as

$$X_{ij} = \begin{cases} +1 & \text{if the observation received the } i^{th} \text{ level of factor } j \\ 0 & \text{otherwise} \end{cases}$$

for $i=1,\ldots, k_j - 1$, and $j=1,\ldots,J$. If interaction effects are present, a simpler correspondence between the ANOVA and regression parameters can be made if the indicator variables are defined as

$$X_{ij} = \begin{cases} +1 & \text{if the observation received the } i^{th} \text{ level of factor } j \\ -1 & \text{if the observation received the } k_j{}^{th} \text{ level of factor } j \\ 0 & \text{otherwise} \end{cases}$$

for $i=1,\ldots, k_j - 1$, and $j=1,\ldots,J$. The interaction terms are constructed by multiplication of the appropriate main effect indicator variables.

Once the correspondence between the ANOVA parameters and the regression parameters has been established, hypotheses concerning the ANOVA parameters can be tested with the regression model using the general linear test approach (see, for example, Neter and Wasserman [1974]).

3. The Multivariate Relationship

The generalization of the relationship between regression analysis and ANOVA to the multivariate case can be illustrated in the same manner as described in the previous section. This approach is not shown in many applied multivariate textbooks but is appealing because it not only provides the student with a general framework for relating the techniques learned in the multivariate course, but it also relates the multivariate techniques to the univariate statistics learned in previous courses. Through the use of indicator variables, multivariate analysis of variance (MANOVA) hypotheses can be formulated in terms of the parameters of a multivariate regression model and tested using a likelihood ratio test. This procedure follows the same general framework discussed in the previous section.

In particular, consider a balanced randomized complete block design with k blocks, r treatments, and p response variables. The MANOVA model is represented as

$$Y_{ij\alpha} = \mu_\alpha + \tau_{i\alpha} + \beta_{j\alpha} + \varepsilon_{ij\alpha}; \quad \begin{matrix} i=1,\ldots,r \\ j=1,\ldots,k \\ \alpha=1,\ldots,p \end{matrix} \qquad (3.1)$$

where $Y_{ij\alpha}$ represents the observation on the α^{th} response variable for the i^{th} treatment in the j^{th} block, μ_α is the overall mean of the α^{th} response variable, $\tau_{i\alpha}$ is the i^{th} treatment effect associated with the α^{th} response variable, $\beta_{j\alpha}$ is the j^{th} blocking effect associated with the α^{th} response variable, and $\varepsilon_{ij\alpha}$ are normally distributed random errors with zero means, variances $\sigma_{\alpha\alpha}$ and covariance given by

$$\text{Cov}(\varepsilon_{ij\alpha}, \varepsilon_{i'j'\alpha'}) = \sigma_{\alpha\alpha'} \text{ if } i=i',\ j=j' \qquad (3.2)$$

$$= 0 \text{ otherwise}$$

A multivariate regression model using indicator variables as defined in Section 2 is written as

$$Y = X\Gamma + E \qquad (3.3)$$

where Y is a (N x p) matrix of the observed response variables, N is the total number of observations, Γ is a ((k + r - 1) x p) matrix of the form

$$\Gamma = \begin{matrix} 1 \\ \\ (k-1) \\ \\ (r-1) \end{matrix} \begin{bmatrix} \gamma_{01} & \cdots & \gamma_{0p} \\ \hline \gamma_{b11} & \cdots & \gamma_{b1p} \\ \vdots & & \\ \gamma_{b(k-1)1} & \cdots & \gamma_{b(k-1)p} \\ \hline \gamma_{t11} & \cdots & \gamma_{t1p} \\ \vdots & & \\ \gamma_{t(r-1)1} & \cdots & \gamma_{t(r-1)p} \end{bmatrix} = \begin{bmatrix} \gamma_0 \\ \gamma_b \\ \gamma_t \end{bmatrix} \qquad (3.4)$$

X is a (N x (k + r - 1)) matrix of the form

$X = \underline{x}_\ell'$, $\ell = 1,\ldots,N$,

$\underline{x}_\ell' = [1\ x_{\ell b1} \cdots x_{\ell b(k-1)}\ x_{\ell t1} \cdots x_{\ell t(r-1)}]$,

$$x_{\ell bh} = \begin{cases} 1 \text{ if observation } \ell \text{ is in block } h, \\ 0 \text{ otherwise} \end{cases} \qquad (3.5)$$

$h = 1,\ldots,(k-1)$,

$$x_{\ell tm} = \begin{cases} 1 \text{ if observation } \ell \text{ received treatment } m, \\ 0 \text{ otherwise} \end{cases}$$

$m = 1,\ldots,(r-1)$.

and E is a (N x p) matrix of error terms.

The test for treatment effects in terms of the MANOVA model given in (3.1) is

$$H_o: \begin{bmatrix} \tau_{11} \\ \vdots \\ \tau_{1p} \end{bmatrix} = \cdots = \begin{bmatrix} \tau_{r1} \\ \vdots \\ \tau_{rp} \end{bmatrix} \qquad (3.6)$$

Given the definition of the indicator variables shown in (3.5), (3.6) is true if $\gamma_t = 0$ where γ_t is defined in (3.4). Kshirsagar [1972] shows the likelihood ratio criterion for testing the hypothesis $\gamma_t = 0$ can be expressed as

$$\Lambda_H = \frac{|A|}{|A + B - B_t|} \qquad (3.7)$$

where B is the regression sums of squares and crossproducts matrix for the regression of the vector of response variables on the vector of the entire set of indicator variables, A is the residual sums of squares and crossproducts matrix for the same regression, and B_t is the regression sums of squares and crossproducts matrix for the regression of the vector of response variables on the vector of indicator variables omitting the variables x_{t1} through $x_{t(r-1)}$. This test criterion can be calculated using an approach similar to the univariate general linear test and a canonical correlation analysis software program.

Canonical analysis is the extension of regression of one variable on several variables to the regression of several variables on several variables. Kshirsagar [1972] provides an excellent discussion of the relationships among canonical analysis, MANOVA, and discriminant analysis and provides a strong case for the statement that canonical analysis is the most general type of linear model. Canonical analysis can be used to calculate Λ_H in the following manner. First, a canonical analysis is performed with the vector of response variables as the first variable set and the entire vector of indicator variables as the second set. This model is referred to as the full model. The associated value of Wilk's Λ is $\Lambda_{FULL} = |A|/|A+B|$ where A and B are as defined previously. This value can be expressed in terms of the squares of the sample canonical correlation coefficients (eigenvalues) as $\Lambda_{FULL} = (1 - \theta_1)(1 - \theta_2) \ldots (1 - \theta_q)$ where $q = \min(p, k + r - 2)$ and $\theta_1, \theta_2, \ldots, \theta_q$ are the eigenvalues of the canonical analysis (see, for example, Kshirsagar [1972, p. 308]).

The second step is to perform a canonical analysis for the reduced model where $\gamma_t = 0$. This is accomplished by again using the vector of response variables as the first set but for the second set taking the set of indicator variables omitting the x_{tm}'s. This reduced model will yield $\Lambda_{RED} = |A+B-B_t|/|A+B| = (1-\theta_1^*)(1-\theta_2^*) \ldots (1-\theta_s^*)$ where A, B, and B_t are as defined previously, $s = \min(p, k-1)$, and $\theta_1^*, \theta_2^*, \ldots, \theta_s^*$ are the eigenvalues associated with the canonical analysis of the reduced model. It can now be seen that the appropriate test statistic given in (3.7) can be represented as $\Lambda_H = \Lambda_{FULL}/\Lambda_{RED}$. The calculated value of Λ_H can then be used in Rao's approximate F or Bartlett's approximate χ^2 statistic to test H_o: $\gamma_t = 0$.

4. An Example

The data shown in Table 1 was obtained from blood samples drawn from patients that were used in an experiment to compare four different brands of a reagent that is mixed with blood in order to perform a chemical analysis.[1] Brand A is the reagent presently being used in the laboratory and Brands B, C, and D are three different batches of a new, less expensive reagent. The question of interest is whether there exists a difference among the four reagents in regard to the measurements that are recorded for several blood properties. The properties reported in Table 1 are white blood count, red blood count, and hemoglobin content.

Each blood sample drawn from a patient was first separated into four parts and then one of the four reagents was assigned at random to each part. The patients can be treated as blocks and the

TABLE 1
Blood Data

Subject	Reagent A	Reagent B	Reagent C	Reagent D
1	(8.0, 3.96, 12.5)*	(8.0, 3.93, 12.7)	(7.9, 3.86, 13.0)	(7.9, 3.87, 13.2)
2	(4.0, 5.37, 16.9)	(4.2, 5.35, 17.2)	(4.1, 5.39, 17.2)	(4.0, 5.35, 17.3)
3	(6.3, 5.47, 17.1)	(6.3, 5.39, 17.5)	(6.0, 5.39, 17.2)	(6.1, 5.41, 17.4)
4	(9.4, 5.16, 16.2)	(9.4, 5.16, 16.7)	(9.4, 5.17, 16.7)	(9.1, 5.16, 16.7)
5	(8.2, 5.16, 17.0)	(8.0, 5.13, 17.5)	(8.1, 5.10, 17.4)	(7.8, 5.12, 17.5)
6	(11.0, 4.67, 14.3)	(10.7, 4.60, 14.7)	(10.6, 4.52, 14.6)	(10.5, 4.58, 14.7)
7	(6.8, 5.20, 16.2)	(6.8, 5.16, 16.7)	(6.9, 5.13, 16.8)	(6.7, 5.19, 16.8)
8	(9.0, 4.65, 14.7)	(9.0, 4.57, 15.0)	(8.9, 4.58, 15.0)	(8.6, 4.55, 15.1)
9	(6.1, 5.22, 16.3)	(6.0, 5.16, 16.9)	(6.1, 5.14, 16.9)	(6.0, 5.21, 16.9)
10	(6.4, 5.13, 15.9)	(6.4, 5.11, 16.4)	(6.4, 5.11, 16.4)	(6.3, 5.07, 16.3)
11	(5.6, 4.47, 13.3)	(5.5, 4.45, 13.6)	(5.3, 4.46, 13.6)	(5.3, 4.44, 13.7)
12	(8.2, 5.22, 16.0)	(8.2, 5.14, 16.5)	(8.0, 5.14, 16.5)	(7.8, 5.16, 16.5)
13	(5.7, 5.10, 14.9)	(5.6, 5.05, 15.3)	(5.5, 5.02, 15.4)	(5.4, 5.05, 15.5)
14	(9.8, 5.25, 16.1)	(9.8, 5.15, 16.6)	(8.1, 5.10, 13.8)	(9.4, 5.16, 16.6)
15	(5.9, 5.28, 15.8)	(5.8, 5.25, 16.4)	(5.7, 5.26, 16.4)	(5.6, 5.29, 16.2)
16	(6.6, 4.65, 12.8)	(6.4, 4.59, 13.2)	(6.3, 4.58, 13.1)	(6.4, 4.57, 13.2)
17	(5.7, 4.42, 14.5)	(5.5, 4.31, 14.9)	(5.5, 4.30, 14.9)	(5.4, 4.32, 14.8)
18	(6.7, 4.38, 13.1)	(6.5, 4.32, 13.4)	(6.5, 4.32, 13.6)	(6.5, 4.31, 13.5)
19	(6.8, 4.67, 15.6)	(6.6, 4.57, 15.8)	(6.5, 4.55, 16.0)	(6.5, 4.56, 15.9)
20	(9.6, 5.64, 17.0)	(9.5, 5.58, 17.5)	(9.3, 5.50, 17.4)	(9.2, 5.46, 17.5)

*The values in parentheses are white blood count in thousands of cells per cubic millimeter, red blood count in millions of cells per cubic millimeter, and hemoglobin in grams per deciliter, respectively.

data analyzed as a multivariate randomized block experiment with $k = 20$, $r = 4$, and $p = 3$. The analysis described in Section 3 is performed using BMDP6M of the Biomedical Computer Programs [1975] and is reported in Table 2.

5. Conclusion

Although the relationship between regression analysis and ANOVA is often illustrated in many intermediate statistical textbooks through the use of indicator variables, many applied textbooks in multivariate analysis do not satisfactorily discuss the generalization of this relationship. This paper makes note of this relationship and shows how it can be presented, drawing from material learned in previous statistical courses.

FOOTNOTES

[1]I wish to thank Brad Hill, M.A., M.T. (ASCP), Administrative/Chief Technologist of the Denver Clinic for the use of this data.

REFERENCES

Dixon, W. J., ed. [1975], BMDP: Biomedical Computer Programs, Los Angeles: University of California Press.

Green, Paul E. [1978], Analyzing Multivariate Data, Hinsdale, Illinois: The Dryden Press.

Kshirsagar, Anant M. [1972], Multivariate Analysis, New York: Marcel Decker, Inc.

Neter, John and Wasserman, William [1974], Applied Linear Statistical Models, Homewood, Illinois: Richard D. Irwin, Inc.

Wesolowsky, George O. [1976], Multiple Regression and Analysis of Variance, New York: John Wiley and Sons, Inc.

TABLE 2

Calculation of Λ_H For Blood Data Using BMDP6M

Full Model

$$\theta_1 = .99880$$

$$\theta_2 = .99508$$

$$\theta_3 = .86850$$

$$\Lambda_{FULL} = (1 - .99880)(1 - .99508)(1 - .86850)$$

$$= 7.7638 \times 10^{-7}$$

Reduced Model

$$\theta_1^* = .99818$$

$$\theta_2^* = .98340$$

$$\theta_3^* = .75398$$

$$\Lambda_{RED} = (1 - .99818)(1 - .98340)(1 - .75398)$$

$$= 7.4328 \times 10^{-6}$$

$$\Lambda_H = \Lambda_{FULL}/\Lambda_{RED}$$

$$= .1045$$

AN EMPIRICAL APPROACH TO TEACHING GENERAL STATISTICS AUGMENTED BY THE COMPUTER

Robert J, Oberg, Framingham State College

1. Introduction

Teaching a course in elementary statistics is quite challenging. Among the reasons are

-- it is frequently a required course, resulting in a captive audience

-- students may come to the course with a weak background in mathematics

-- it is a completely new subject to most students; they very likely have had little previous exposure to statistics.

To the instructor the biggest challenge may be how to teach concepts and understanding of this novel subject to students who do not have the background to understand the underlying mathematics. It is surely not satisfactory to follow a "cookbook" approach. That is bad for the student and runs strongly against the grain of an instructor with mathematical training.

One approach is to tie the course closely to the particular field of application of interest to the student (psychological statistics, biostatistics, etc.). When well done this can be a good solution -- strong contact is made with the discipline and the student can see the interpretation and meaning of statistical results. In the author's view, however, such a subject-centered approach may be more appropriate to an advanced course rather than a beginning course. The fundamentals of statistics are really the same no matter what the eventual application may be. In this paper we are concerned with a first course in general statistics.

Another approach is to attempt a mathematical treatment within the constraints imposed by a limited mathematical background on the part of the students. This kind of course will normally include considerable emphasis on non-calculus probability theory, some care in stating assumptions precisely, etc. But proofs really cannot be given, and students may find the mathematical emphasis makes the material too abstract without giving them a true conceptual grasp of statistics.

In this paper we describe a third approach. We frankly admit that given the mathematical background of our students we cannot give mathematical proofs or derivations. But we do not abandon the goal of "proving" our methods. Instead we present experimental evidence for the validity of the laws and methods of statistics. This empirical approach to teaching general statistics will be described in the next section. The following section gives an account of several empirical demonstrations which could be incorporated into a standard course. We conclude with a brief description of computer software to help in implementing the empirical approach.

2. An Empirically Based General Statistics Course

In this section we describe the "pure form" of the empirical approach to teaching statistics. It is based on the next STATLAB by Hodges, Krech and Crutchfield (1975), the only book of its kind of which the author is aware. What will be described below is a model of a course abstracted from STATLAB. The heart of such a course is a large Data Base of real data. In STATLAB the Data Base consists of 32 psychological, physiological and socio-economic variables available for 1296 families. In a future version of his own course the author is contemplating using as the Data Base a subset (of both variables and cases) of the General Social Surveys. For reasons discussed below, the author feels it is definitely advantageous to use real data.

The Data Base provides motivation for the entire course -- here is this huge mass of data: how are we to make sense of it? Students quickly see the need for working with a sample rather than with the entire population. Thus right from the beginning students are aware, in a very concrete way, of the issue of making inferences from a sample to a population, and this remains a theme throughout the course. Each student draws a random sample and works with this sample in progressively more sophisticated ways during the course. The student's homework is not textbook exercises but work with his own random sample from a real population.

The initial analyses are descriptive. Frequency tables and graphs are constructed and summary statistics are calculated. Since each student has his own individual random sample, different students get different answers for sample statistics like $\bar{x}$. This leads to a very natural transition to the study of statistical inference. By pooling the results of members of the class many statistical principles can be convincingly demonstrated. Examples of several such demonstrations are given in the next section.

Some of the benefits of this empirical approach could be achieved with a Data Base of simulated data, but in the author's opinion it is much more advantageous to use real data. Nothing can convey the spirit of statistical inquiry like real data, and students respond in a very positive way to working with actual real-life data, increasing their interest and motivation in the course. Furthermore, the availability of real data opens the door for students to do small research projects of their own choosing. Working with the STATLAB data the author's students have conducted small studies on questions like the effect of mother's smoking on babies' birth weight, the relationship between education and income, and the effect of physical height on the financial success of salesmen (taller men did seem to be more successful!).

Several positive evaluations of the empirical course have been obtained. In Preskenis and Oberg (1979) results of a brief student survey conducted by the author are given. Noteworthy is a median response of 4.80 on a 5 point scale (1 = least favorable and 5 = most favorable) to the question "How did you like the STATLAB approach (working with realistic data) compared to orthodox textbook?" Direct positive feedback has come from students who had been dreading taking this required course

but actually found it to be a positive experience. Favorable comments have been made by professors in other departments (Psychology and Sociology) about their students' having a better grasp of statistical principles after taking the empirical course compared to the orthodox course.

It must be acknowledged that there are definite tradeoffs in the empirical course versus the traditional course. It takes more time to do an extended series of computations with real data than to do textbook exercises, leading to the empirical course possibly covering less material than the standard course. To give an idea of coverage: In a one semester course the author is typically able to cover descriptive statistics, normal distribution, central limit theorem, confidence intervals, hypothesis testing (z-tests, t-test, sign and Wilcoxon texts), and contingency tables. The author has not been able to cover correlation and regression or to introduce analysis of variance.

It should be pointed out that in the author's course considerable emphasis was placed on numerical computation. Although the computer was used extensively in the course to help manage instruction (see Section 4), students were required to do their calculations by hand or with the aid of a pocket calculator. Since many students were quite weak in arithmetic and algebraic skills, it was felt that strengthening these skills through lots of practice was important to their future ability to successfully use statistics as a working tool. If less time in the course were spent on calculation (including computational shortcuts such as coding), more of the standard material could be included within the framework of the empirical course.

3. Selected Empirical Demonstrations

The key to enhancing the student's conceptual understanding of statistical principles through the empirical approach lies in various well chosen demonstrations that can be made at various points in the course. In fact such demonstrations can be incorporated into a standard course provided that a suitable Data Base of preferably real data is available. Each student should be asked to draw his own random sample and perform various calculations which can subsequently be compared to similar calculations made by his classmates.

Distribution of Sample Statistics

A concept that is often difficult for beginning students to grasp is that a sample statistic like $\bar{X}$ is not merely a number that is computed but is also a random variable that has a distribution of its own. Using the empirical approach each student will have calculated a value of $\bar{x}$ for his own random sample. These individual $\bar{x}$'s are pooled by the instructor to form a class distribution for which a frequency tabulation can be made and summary statistics computed. This leads to a convincing proof of the central limit theorem, both regarding the shape of the $\bar{X}$ distrubution and the relationship of the standard deviation of $\bar{X}$ to the standard deviation of X. The properties of the distributions of other sample statistics such as the median and standard deviation are demonstrated in a similar manner.

Confidence Intervals

Students appreaciate the inherent variability of a statistic like $\bar{x}$ from the fact that their classmates all have computed slighly different values. After the concept of confidence interval has been introduced each student can obtain a measure of the precision of his estimate of the population mean by calculating, for example, a 68% confidence interval. At this time the instructor can reveal the true population mean, and by a show of hands it can be seen what proportion of the class will have captured the population mean in their confidence intervals. Hopefully this proportion will be close to 68%. The same demonstration can be repeated for other population parameters.

Effect of Non-Normality

The empirical approach can help foster a more sophisticated attitude on the part of the student with respect to the role of various assumptions, such as normality, on statistical theorems. For example, in the STATLAB course students work very early with two random variables, height and income. The first is almost perfectly normal and the second is highly skewed. Students will find good agreement between theory and practice in the distributions of $\bar{X}$ and S for height but will observe marked discrepancies for income (assuming a blind extrapolation of predictions valid when the underlying variable is normal).

Hypothesis Testing

Students are given two null hypotheses to test. The first is known by the instructor (on the basis of population parameters) to be true and the second is known to be false. Using test statistics computed from their own random samples students carry out several tests of varying power (e.g. t-test, Wilcoxon test, sign test) of each null hypothesis. Afterwards the instructor reveals which hypothesis was really true and which false. A poll can be taken to see how many students made false rejections and false acceptances. This demonstration brings home the conepts of Type I and Type II errors and of the power of a statistical test.

4. Computer Aids

Having each student work with his own individual random sample is at the crux of the empirical approach, but it can create a very practical problem. Since each student gets a different value from his classmates' for each statistic he calculates, how is he to know if he has done his work correctly? It is impractical for an instructor or teaching assistant to handcheck each student's calculations. When students do not get feedback on their work and are not evaluated, many students may lose motivation and wind up not doing the assignments, making class discussion and pooling of results difficult.

The author has developed a package of computer programs to help solve these problems. What these

programs do in brief is check each student's calculations. During the semester students are given various assignments in which they are asked to keypunch some of the statistics they have calculated for their own random samples. For each assignment the programs will print a report for each student showing the correct computations and write to a file the correct answers. This file can be used as input to an SPSS program to tabulate the class distribution. The programs also assign a grade to each student for each assignment and update a cumulative grade file.

The programs are written in structured Fortran. Because of the extensive file manipulations involved the programs at present can only be run on a CDC Cyber computer under the NOS or KRONOS operating systems. Further information about these programs can be found in Preskenis and Oberg (1979).

References

Hodges, J.L. Jr., Krech, David and Crutchfield, Richard S. (1975).

STATLAB: An Empirical Introduction to Statistics, New York: McGraw-Hill Book Co.

Preskenis, Kenneth and Oberg, Robert (1979). "Teaching Statistics and Using Computers". Proceedings of the National Educational Computing Conference.

A PROBABILITY MODEL FOR MULTIPLE CHOICE TESTS

Richard M. Rubison, Kansas State University

1. Introduction

In this paper a Beta-Binomial distribution is proposed for modeling the number of questions answered correctly on a multiple choice test. This model distinguishes between the number of correct answers obtained through knowing the answer and the number of correct answers obtained including correct guesses. The derived formulas extend those of Morrison [1]. Comparisons between two scoring (grading) procedures are made.

2. The Model

Let

- m = the number of questions on a test.
- n = the number of choices on each question.
- p = the probability of knowing the answer to a randomly selected question.
- Π = the probability of giving a correct answer to a randomly selected question.

Thus,

$$\Pi = p+(1-p)(1/n)$$

and

$$p = (n/(n-1))\Pi-(1/(n-1)).$$

We make the assumption that p is distributed in the population as a beta random variable Be(α, β) with density

$$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, p^{\alpha-1}(1-p)^{\beta-1}$$

where $0 \le p \le 1$.

Define the unobservable random variable K to be the number of correct answers obtained by knowing the answer.

Hence,

$$(1)\quad Pr[K=i] = \int_0^1 \binom{m}{i} p^i (1-p)^{m-i} f(p)\,dp$$

$$= \binom{m}{i} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+i)\Gamma(\beta+m-i)}{\Gamma(\alpha+\beta+m)}$$

for i=0, 1, ..., m.

and K has a Beta-Binomial Distribution [2]. Application of Fubini's theorem gives the mean and variance of K respectively as

$$(2)\quad E[K] = m\frac{\alpha}{(\alpha+\beta)}$$

$$(3)\quad V[K] = \frac{m\,\alpha\,\beta}{(\alpha+\beta)^2} \frac{(\alpha+\beta+m)}{(\alpha+\beta+1)}$$

We now define the random variable J to be the number of correct answers. Then,

$$Pr[J=i] = \int_{\frac{1}{n}}^1 \binom{m}{i} \Pi^i (1-\Pi)^{m-i} g(\Pi)\,d\Pi$$

where $g(\Pi) = f((n/(n-1))\Pi-(1/(n-1)))$

with $\frac{1}{n} \le \Pi \le 1$. By a change of variables,

$$(4)\quad Pr[J=i] = \int_0^1 \binom{m}{i} \{((n-1)/n)\,p + (1/n)\}^i \{((n-1)/n)(1-p)\}^{m-i} f(p)\,dp$$

Since, i is always an integer we can write

$$(p + \tfrac{1}{n-1})^i = \sum_{j=0}^{i} \binom{i}{j} p^j (1/(n-1))^{i-j}$$

therefore

$$(5)\quad Pr[J=i] = \binom{m}{i}\left(\frac{n-1}{n}\right)^m \sum_{j=0}^{i} \binom{i}{j} (1/(n-1))^{i-j} \frac{\Gamma(\alpha+j)\Gamma(\beta+m-i)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+j+m-i)\Gamma(\alpha)\Gamma(\beta)}$$

for i = 0, 1, ..., m.

Again, application of Fubini's theorem using equation (2) gives

$$E[J] = m\,\frac{\alpha}{(\alpha+\beta)} + \frac{m}{n}\,\frac{\beta}{(\alpha+\beta)}$$

$$(6)\quad = E[K] + \frac{(m-E[K])}{n}$$

Similarly,

$$V[J] = \left(\frac{n-1}{n}\right)^2 \frac{m\alpha\beta(\alpha+\beta+m)}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{(n-1)}{n}\,\frac{\beta}{(\alpha+\beta)}$$

$$(7)\quad = \left(\frac{n-1}{n}\right)^2 V[K] + \frac{(n-1)}{n^2}\,(m-E[K])$$

3. Two Special Cases

If f(p) is degenerate at p_o then K is distributed as a binomial random variable Bin (m, p_o) while J is distributed as Bin $(m, p_o+\frac{1}{n}(1-p_o))$. In this case, the probability of knowing the answer to a randomly selected question is a fixed constant p_o.

The second special case consists of a population composed of a mixture of known questions and guessed questions. The corresponding distribution f(p) is

concentrated on two points, one and zero, with probabilities p_0 and $(1-p_0)$ respectively.

4. Scoring Alternatives

Let C_1 = the score (weight assigned to each correct answer
and C_2 = the score (weight) assigned to each incorrect answer.
We consider the problem [3] of determining scores C_1 and C_2 such that the expected score has the value E[K].
Since the score

$$S = C_1J+C_2(m-J)$$

$$(8) = (C_1-C_2)J+C_2m$$

then

$$(9)\ E[S] = (C_1-C_2)E[J]+C_2m$$

and the problem consists of finding values C_1 and C_2 satisfying

$$(10)\ E[K] = (C_1-C_2)E[J]+C_2m$$

for all values of α and β. Using (6), the solution is $C_1=1$, $C_2=-(1/(n-1))$, i.e, we assign 1 point for each correct answer and subtract $(1/(n-1))$ for each incorrect answer. With this scoring convention, say

$$S_1 = \frac{n}{(n-1)}J - \frac{m}{(n-1)}$$

we have

$$V[S_1] = (\frac{n}{n-1})^2V[J] = V[K]+\frac{1}{(n-1)}(m-E[K])$$

and since $E[S_1] = E[K]$

$$V[K] = V[S_1] - \frac{1}{(n-1)}(m-E[S_1]).$$

It is interesting to note that under the usual scoring procedure with $C_1=1$, $C_2=0$ then $S_2=J$ and the expected score is the expected number of correct answers and not the expected number of correct answers obtained through knowing.

References

[1] Morrison, D. G. (1978), "A Probability Model for Forced Binary Choices" The American Statistician, 32: 1, 23-25.

[2] Johnson, N. L. and Kotz, S. (1969), Discrete Distributions, New York: John Wiley and Sons, Inc.

[3] Austin, J. D. (1978), "Grading Answer-Until-Correct Tests", The American Mathematical Monthly, 85:7, 588-590.

A TEACHER'S AID FOR THE BINOMIAL PROBABILITY DISTRIBUTION: A COLLEGE FOOTBALL CONTEST

Ronald E. Shiffler, Georgia State University
James R. Lackritz, San Diego State University

ABSTRACT

During the 1978 college football season, the Atlanta Constitution ran a weekly football contest consisting of fifteen randomly selected games. Anyone predicting the correct outcomes of at least fourteen games was awarded a tee-shirt. The contest provides examples in probability, estimation and hypothesis testing. The theoretical probability of winning a tee-shirt is solved with the binomial distribution. The contest results are used to estimate this probability. A test of hypothesis is conducted to check guessing versus expertise. Bounds on the number of free tee-shirts to be awarded in future contests of similar size are formed. Finally, a nonparametric test to detect improvement in prediction over time is performed.

INTRODUCTION

Many topics typically taught in introductory statistics courses are difficult to illustrate with "real-life" examples, especially ones that appeal to the majority of the class. Occasionally, newspapers and/or magazines publish stimulating examples that are adaptable to the classroom. Our daily routine of scouring the sports pages turned up an interesting application of the binomial probability distribution via college football games.

During the 1978 college football season, the Atlanta Constitution ran a weekly football contest, dubbed "Pigskin Picks", consisting of fifteen randomly selected games. The contest began with games of September 9th and ended with games of November 25th, a total of twelve weeks. Each person entering the contest was required to select the correct outcome (win or tie) of each of the fifteen games. No one was permitted to enter more than once per week, but everyone was eligible again the next week. One weekly winner was crowned based on the greatest number of correct picks and a "tiebreaker" scheme, if necessary. In addition, anyone picking fourteen or fifteen games correctly was awarded a Super Picker tee-shirt.

For teachers of introductory statistics, we discuss four applications of Pigskin Picks: (1) probability, (2) estimation, (3) hypothesis testing, and (4) expectation. First let us reproduce the data that we collected from the results published in the Atlanta Constitution every Tuesday after the Saturday games. For $i = 1, 2, \ldots, 12$, let

w_i = week number i

n_i = the number of entries for week i

t_i = the number of entries for week i with fourteen or more correct selections

t_i/n_i = the proportion of entrants for week i receiving a free tee-shirt

$\hat{p}_i$ = an estimate of p, the probability of correct selection for an individual game during week i (the rank of $\hat{p}_i$ is in parenthesis)

TABLE 1

w_i	Date of games	n_i	t_i	t_i/n_i	$\hat{p}_i$
1	Sept. 9	4602	2	.00043459	.4956 (4)
2	Sept. 16	4582	1	.00021825	.4704 (2)
3	Sept. 23	4512	2	.00044326	.4964 (6)
4	Sept. 30	5119	8	.00156281	.5466 (7)
5	Oct. 7	4736	207	.04370777	.7126 (11)
6	Oct. 14	5068	0	.00000000	.0000 (1)
7	Oct. 21	4959	49	.00988102	.6315 (8)
8	Oct. 28	5202	71	.01364860	.6480 (10)
9	Nov. 4	5123	759	.14815538	.7914 (12)
10	Nov. 11	4733	61	.01288823	.6450 (9)
11	Nov. 18	4513	2	.00044316	.4963 (5)
12	Nov. 25	4174	1	.00023958	.4737 (3)
		57323	1163		

PROBABILITY APPLICATION

Suppose we wish to determine the probability of winning a tee-shirt. Let X represent the number of correct selections out of fifteen games. Then X satisfies the characteristics of a binomial random variable [1,p.129] with $n = 15$. We seek $P(X \geq 14)$. However, this probability depends on the value of p, the probability of making a correct selection. (It may be helpful to the students to define p as the percentage of correct picks over the course of a season.) "Experts" may boast of 80% accuracy in predicting while others may profess to have no knowledge of football in which case coin tossing or random guessing can be used to simulate selection. Using both extremes and some intermediate values, we generate the following probabilities of winning a tee-shirt.

TABLE 2

p	$P(X \geq 14)$
.50	.000488
.55	.001692
.60	.005172
.65	.014179
.70	.035268
.75	.080181
.80	.167126

What do these numbers mean? For $p = .5$ we interpret $P(X \geq 14)$ thusly: for every group of 10,000 entries in which each entry is marked randomly, an average of (approximately) 5 tee-shirts will be awarded.

ESTIMATION APPLICATION

Define a binomial random variable T to be the number of entrants receiving a free tee-shirt.

If we treat the entire data set as a sample of $N = 57,323$ entries, then an estimator of $\theta = P(X \geq 14)$ is $\hat{\theta} = T/N$ which possesses an approximately normal distribution with mean and variance equal to θ/N and $\theta(1-\theta)/N$, respectively. One can verify that $\hat{\theta} \pm 3\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{N}}$ is a subset of $[0,1]$, hence an approximate 95% confidence interval for θ is

$$\hat{\theta} \pm 1.96\sqrt{\frac{\theta(1-\theta)}{N}}$$

$$\approx \frac{1163}{57323} \pm 1.96\sqrt{\frac{(.020288)(.979712)}{57323}}$$

$$= .020288 \pm .001154$$

Assuming the 1978 season is a random sample[1] from the population of all college football seasons, then we are approximately 95% confident that between 1.9134% and 2.1442% of all entrants will receive a free tee-shirt.

Additionally, we place bounds [2,p.227] on the number of free tee-shirts awarded per season. Such information may be important to the newspaper when ordering tee-shirts. The mean and variance of T are $N\theta$ and $N\theta(1-\theta)$, respectively. Upper and lower limits for T are

$$N\theta \pm 2\sqrt{N\theta(1-\theta)}$$

$$\approx 1163 \pm 2\sqrt{(57323)(.020288)(.979712)}$$

$$= 1163 \pm 67.5$$

Thus we anticipate awarding between 1095 and 1230 free tee-shirts per season.

HYPOTHESIS TESTING APPLICATION

In general, are entries filled out by (essentially) guessing or with some football "expertise"? One approach to this question involves testing hypothesis about p:

H_0: $p = .5$
H_a: $p > .5$

Since $p = .5$ if and only if $\theta = P(X \geq 14) = .000488$, we can equivalently test

H_0: $\theta = .000488$
H_a: $\theta > .000488$
Test statistic:

$$z = \frac{\hat{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{N}}} = \frac{.020288 - .000488}{\sqrt{\frac{(.000488)(.999512)}{57323}}}$$

$$= 214.65$$

Rejection region: Reject H_0 if $z > 1.645$ ($\alpha = .05$)

Conclusion: Reject H_0 with $\alpha = .05$

There is sufficient evidence to conclude that most people entering the Pigskin Picks contest do so with some knowledge of football.

Perhaps this knowledge "grows" as the season progresses, making the last few weeks of the season easier to predict. We check for such a phenomenon by computing Spearman's correlation coefficient [1,p.420], r_S, between w_i and the ranked (from low to high) values of $\hat{p}_i$. Letting $d_i = w_i - \text{rank}(\hat{p}_i)$, we have

$$r_S = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = .231$$

Is this sample correlation sufficient evidence to conclude a positive correlation in the population?

H_0: $\rho_S = 0$
H_a: $\rho_S > 0$
Test statistic: $r_S = .231$
Rejection region: Reject H_0 if $r_S > .499$ ($\alpha = .05$)
Conclusion: Do not reject H_0

No, the data are insufficient to support the theory that games during the latter weeks of the season are easier to predict.

EXPECTATION APPLICATION

Although the contest is usually played for enjoyment only, a true probabilist is interested in his/her expected gain. Suppose a tee-shirt costs (is worth) \$5 and it costs \$.15 to mail-in the entry each week. Then, the weekly expected gain is

$$E(\text{gain}) = \theta \cdot \$5 + (1-\theta) \cdot (-\$.15)$$

What value must θ assume to make this a "fair contest"? When $\theta = .029126$ (or equivalently when $p = .689$), $E(\text{gain}) = 0$. Pigskin Picks is therefore a "losing" proposition whenever one's predictive ability is below 68.9% for a given week.

CONCLUSION

To stimulate interest in a few of the major topics in most introductory statistics courses - probability, estimation, hypothesis testing, and expectation - a weekly college football contest provides many possibilities for examples, discussion, test questions, and/or case studies.

REFERENCES

[1] McClave, James T. and Frank H. Dietrich, II, (1979): Statistics, Dellen Publishing Company, San Francisco, California.

[2] Mendenhall, William, (1979): Introduction to Probability and Statistics, 5th. Ed., Duxbury Press, North Scituate, Massachusetts.

[1] This contest has been running at least two years although we have collected data for the 1978 season only.

A TASK-ANALYSIS APPROACH TO DESIGNING A REGRESSION ANALYSIS COURSE

Donald G. Watts, Queen's University

1. Introduction

My approach to education has changed markedly in the last few years, primarily due to the influence of two authors: Robert F. Mager, author and co-author of several sensible simple books on instructional objectives and other aspects of course development (2,3,4,5,6) and Ivor K. Davies who emphasizes the resource-management aspects of education (1). Another excellent reference, concerning teaching styles and effectiveness, is the extremely helpful and well-written manual by Saunders, Welsh and Hansen(7).

A common feature in all these references is the need to define carefully, in an operational way, the course to be taught. One excellent and thorough approach to this detailed in Mager and Beach (6), begins by writing a *job description*, which is "a general statement about what a person on the job does, and tells something about the conditions under which he does them" (p.7). Next is the *task analysis*, which requires "listing *all* the tasks that might be included in the job" (p.10) and then listing "the steps involved in each of the tasks on the list in terms of what the person *does* when performing the step" (p.14). One then tries to identify these tasks and steps which the target population, or incoming class, can not already do so as to be able to develop the *instructional objectives* for the course.

Instructional objectives are clear statements of "the kind of performance that will be expected at the end of the course" (p.29), and as such can provide a detailed and extensive description of the course. They are great aids to an instructor in clarifying course content and in organizing a course, especially as regards developing reinforcing assignments. For further details, I recommend all of the above references.

As emphasized by Mager and Beach (p.29) "The key question to ask is this: What kinds of things should the student be able to do at the end of the course that will most facilitate his becoming a skilled craftsman in the least of time?" Now, the word "craftsman", and indeed the job description - task analysis - instructional objectives approach is often *extremely* disturbing to academics. One must remember that most of Mager's work deals with vocational instruction, and as such can not be transported indiscrimately into a more intellectual area: nevertheless, careful exploitation of the ideas, techniques and philosophy can be very beneficial to course and program development. To illustrate use of the approach in statistics, I offer a description of an early iteration in the process of revising an undergraduate course in least squares regression analysis.

2. Developing objectives for an undergraduate regression analysis course.

2.1 Background.

The course in question is a third year undergraduate course which has, as prerequisites, a one semester course in statistics, about two semesters of calculus, and one semester of linear algebra. In fact, most of the students have an adequate exposure to confidence intervals and t tests, ordinary derivatives and simple matrix operations.

2.2 A job description for a regression analyst.

I do not suggest that there ever should or would be a person who would be solely described as a regression analyst, but *for purposes of course development* I have used the following:

JOB DESCRIPTION FOR A REGRESSION ANALYST

Regression analysis can be described as a process by which a descriptive or predictive relationship is derived from data consisting of values of a response variable, y, and corresponding values of the predictor or regressor variables, $x_1, x_2 \ldots$.

A regression analyst must be able to recognize when regression analysis is appropriate and, when it is, be able to perform the analysis. A proper analysis will include knowing when the assumptions necessary for least squares regression are appropriate and if they are not, how to modify the analysis by transforming or doing weighted least squares. The analyst may also have to devise an appropriate linear, nonlinear, or spline model or help a client do so.

Having selected the form of the model or decided to use a computer-aided approach, the analyst must be able to use computer programs to estimate the parameters in the model, to examine and plot residuals, and to interpret the program output (especially for the client, if one is involved) so as to permit improvement of the model and to provide information for future experiments to validate or repudiate the model. The analyst must be able to describe the analysis and the findings in a clear, concise meaningful report. If a client is involved, the report must be written in terms that the client understands and appreciates.

2.3 Task Analysis.

Using the above job description, an extensive list of tasks was developed, and then detailed. The task analysis covers the performance of a highly-skilled regression analyst, all the tasks which the regression analyst must perform, and especially those which must be mastered. Each task was written out and classified according to frequency of performance, importance, and learning difficulty, and then a detailed task listing developed for each task. These items were categorized according to type of performance (e.g. difcriminate, recall, calculate, plot, etc.) and learning difficulty.

In this instance, I developed nine tasks, as follows:

Task list for a regression analyst.

1. Understand the problem and the observational or experimental situation,
2. Decide if regression analysis is appropriate,
3. Perform a preliminary analysis,
4. Transform to stabilize variance or to simplify the model,
5. Select an appropriate form of model,
6. Estimate the parameters,
7. Refine the model,
8. Evaluate the model,
9. Write a report

The task details followed. For the first task, for example, I had six details, as:

1.1 Read a report or study the situation and then write a brief concise description of the problem, placing emphasis on why the problem is important,

1.2 Discuss this description with the client to ensure that you have a proper understanding of the situation,

1.3 Carefully identify the variables involved as responses or predictors,

1.4 Note any distinctive features of the data or the situation which may be helpful in choosing transformations of any of the variables or in elucidating the nature of inter-relationships,

1.5 Visit the experimental site or carefully examine how the data are collected to clarify any points and to better understand the situation. Ask questions!

1.6 Determine what the important questions are. Is the main aim to estimate parameters, to test whether a subset of the parameters may be zero, to develop a descriptive relationship, to predict future values, or to make decisions?

All told, from the 9 tasks, I generated 56 detailed tasks which, in conjunction with the prerequisites and expected audience, were then used to develop the course objectives.

2.4 Development of objectives

The detailed tasks formed a rich source from which the course instructional objectives were distilled and developed. And although I have not yet managed to write instructional objectives as specifically as recommended in the references concerning conditions under which the objectives will be tested and exactly what level of competence is required, I still find the objectives helpful in delineating a course and especially in setting up assignments. For this course based on the 9 tasks and 56 detailed tasks, a total of 48 course objectives were developed and categorized as shown in the Appendix.

3. Experience with this approach.

As can be imagined, planning a course in this manner entails a good deal of work by the instructor — and as also can be imagined, that work must be done well in advance of the course. I estimate that I spent upwards of 200 hours developing these objectives and a detailed coordinated set of assignments to go with them. The investment pays off, however, since more classroom time can be spent discussing student's learning problems than in lecturing. Motivation tends to be high.

More emphasis is naturally placed on student learning rather than professor teaching. Most of the assignments I gave involved real data from Science and other journals, but two simulation experiments were also used. The first was a simple one-factor linear regression situation and the second used a nonlinear model. In both of these cases, each student gets a model whose characteristics and parameters depend on the student's name, so that all the students enjoy a similar challenge. Groups of three or four students also had to find and analyze a real data set. This gives the students an opportunity to run into some of the real-world problems in data sets, and to help someone on campus. This project also develops high motivation to learn and to use regression analysis.

Assignments developed from the course objectives can be effectively organized and graduated for review and motivation. Exercises which build on previous assignments with the same data provide excellent opportunities for review and reinforcement. Small, carefully coordinated steps which allow students to realize their own progress is highly beneficial — the catch is that it is important to mark the assignments carefully, constructively and promptly. This throws a very large burden on the instructor but, given the proper attitude, even this can be a most rewarding responsibility. It is especially valuable in helping the students to master the report-writing skills specified in the last section of the objectives.

By having a detailed set of object-

ives, one can keep very close track on progress through the term and at the end of the term. This is very helpful in deciding what objectives were actually covered and hence how the course needs to be modified. As one's experience changes, both with the material and with classes, so can the objectives.

The job description, task analysis approach to developing instructional obejectives seems to be particularly suited to statistics courses because of the inherent applied nature of statistics. For example, I have found it most helpful in designing a descrete time series analysis course and a spectrum analysis course. However, the approach could also be used to advantage in introductory statistics courses and in theory courses, I believe. Try it, you might like it!

Additional information for the job description and task analysis and objectives should be obtained from other people experienced in the area. In this case I was grateful for assistance and comments from my colleagues at Queen's and from Ron Snee, the E. I. Dupont Company, Willington, Delaware. I would also be grateful for comments and suggestions from readers on both the job description and objectives, so that I may improve the course by basing it on a wider collective experience.

References

Davies, Ivor K., "Competency Based Learning: Technology, Management and Design", McGraw-Hill, New York, 1973.

Mager, Robert F., "Preparing Instructional Objectives", 2d. ed., Fearon Publ., Belmont, Ca., 1975.

_____, "Developing Attitude Toward Learning", Fearon Publ., Belmont, Ca., 1968.

_____, "Measuring Instructional Intent", Fearon Publ., Belmont, Ca., 1973.

_____, "Goal Analysis", Fearon Publ., Belmont, Ca., 1972.

_____ and Beach, Kenneth M. Jr., "Developing Vocational Instruction", Fearon Publ., Belmont, Ca., 1967.

Saunders, Phillip, Welsh, Arthur L., and Hansen, W. Lee, "Resource Manual for Teacher Training Programs in Economics, Joint Council on Economic Education, New York, 1978.

Appendix: Regression analysis course objectives

Competent performance of the tasks described in the job description for a regression analyst requires that the analyst have a good grounding in the theory of regression analysis and experience in applying regression analysis. The objectives for this course therefore include items specifically concerned with theory, under the heading "Foundations": more practical material occurs under the headings "Basics", "Preliminary analysis", "Parameter estimation", "Model refinement, evaluation and validation", and "Report Writing."

I Foundations

Upon completion of Statistics 363, a student is expected to be able to:

1. state the assumptions involved in a least squares regression analysis, and explain their practical importance,
2. derive expressions for variance-stabilizing transformations,
3. state expressions for least squares estimates and state the properties of least squares estimators of the parameters, of the response function and of predictors of future observations, and explain their practical importance,
4. explain analysis of variance as it relates to linear or nonlinear regression,
5. explain iterative nonlinear least squares, describe Marquardt's compromise, and relate it to Gaussian least squares and the method of steepest descent,
6. for nonlinear models state approximate properties of least squares estimators of the parameters, of the response function, and of predictors of future observations, and explain their practical importance.

II Basics

A student is expected to be able to:

1. given a simple model and data set, derive and calculate least squares estimates for a linear model or the increment for a nonlinear model, F1,3,5*
2. given a simple linear model and the least squares estimates, calculate s^2, $X'X$ and $(X'X)^{-1}$ and the parameter correlation matrix. Calculate joint and marginal confidence regions for the parameters and confidence regions for the response function and predicted observations at specified design points, and calculate design and predictive efficiencies and effective sample sizes for the design,
3. given a simple nonlinear model and the least squares estimates, calculate s^2, the derivative matrix V., and approximate confidence regions for the parameters and for the response function and predicted observations at specified points. Be able to use sum of squares contours to delineate approximate confidence regions. F1,3,4,6

III Pre-analysis

Given a written or verbal description of an experimental or observational situation, and the data, a student will be able to:

1. state which assumptions necessary for least squares regression are likely to be appropriate, which are likely to be inappropriate, and which are uncertain, F1
2. decide whether regression analysis is appropriate by correctly summarizing the experiment and identifying the response and predictor variables, F1
3. decide whether there are "errors in both variables," F1
4. decide what type of regression analysis is appropriate, such as ordinary linear least squares, nonlinear least squares, probit analysis, contingency table analysis, weighted least squares or multi-response regression, F1
5. state any distinctive features which suggest a form of model, a transformation to stabilize variance, or transformation of the predictor variables, and correctly specify the design matrix, F1,2
6. state the most important issue to be resolved: that is, is the main aim to estimate parameters, to test for a relationship, to develop a descriptive relationship or a predictive relationship, to predict future observations, or to make decisions, F1,3,4,6
7. decide how to fracture the data to permit model validation, F1,3
8. decide which are the most likely influential predictor variables and state how the response is likely to change with changes in each predictor variable. Also be able to state which predictor variables will likely be involved in interactions.

*indicates which fundamental objectives are involved

IV Preliminary analysis

Given a written or verbal description of an experimental or observational situation and the data, a student will be able to:

1. identify the replication values, if present, and calculate $\bar{y}$ and s^2 for each set,
2. plot the replication deviations from the means on normal probability paper and decide whether the normal assumption is appropriate, and identify any outliers, F1
3. derive 5-number summaries and box-plots for each set of replications. From the box plots decide whether the replication data sets are symmetric and possibly normal. Identify any detached values. F1
4. plot s or s^2 vs $\bar{y}$ and decide whether there is constant variance. If not, use the plot and the theory of variance-stabilizing transformations to derive an appropriate transformation or use the ladder of re-expression to induce symmetry (and hopefully, constant variance),
5. if there are no replications, sketch in a rough fit and calculate residuals. Plot the residuals vs y to help decide if there is constant variance,
6. if it is not possible to use a transformation to stabilize variance, decide what form of weighted least squares would be appropriate, F1,3
7. use simple, multiple or contour plots to derive probable forms of relationships between y and x's, and between transformed y's and transformed x's — without regard to stable variance — to deduce simple relationships, and hence derive a sensible linear or nonlinear model to try,
8. use computer programs to calculate correlation matrices for variables and use the output to select important predictor variables and to detect tendencies for collinearity, F1,3,6
9. use the information obtained or deduced from the previous stages to specify an appropriate first model for fitting, F1

V Parameter estimation

Given a set of data for which a linear model is appropriate, a student will be able to:

1. use a package such as LINREG or MINITAB or SPSS to estimate the parameters in the model, or use a step-wise regression package,

Given a set of data for which a nonlinear model is appropriate, and for which the form of model is specified, a student will be able to:

2. calculate derivatives with respect to the parameters and code the model and the derivatives for use in a nonlinear regression program. If the model has restrictions on any of the parameters, choose a parameterization which will satisfy the restriction and permit estimation of the parameter, F5
3. determine reasonable starting values for the parameters by relating salient features of the response (such as asymptotes, slopes or half-value points) to the parameters, or by plotting the data and deducing starting values, or by linearizing the model and using linear least squares to derive starting values for the original nonlinear parameters,
4. Choose an advantageous parameterization to ensure or speed convergence,
5. use a computer program such as MINITAB or NOTLIN to obtain the least squares estimates.

VI Model refinement, evaluation and validation

Given the output from a linear or nonlinear regression, plus the preliminary analysis, a student will be able to:

1. examine residuals from the fitted model by plotting them against various variables such as run order, $\hat{y}$ and x's , so as to detect any model inadequacies, outliers, non-constant variance or correlations of residuals, F1,3,6
2. make normal plots to detect outliers or inappropriateness of the assumption of normal errors, F1
3. if there are possible outliers, either eliminate the observations after having satisfactorily justified such action and re-estimate the parameters, or use robust regression, or incorporate new variables (such as interactions) to account for the behaviour, F1,3,5
4. compute correlations of residuals with other potential predictor variables to detect additional important variables,
5. interpret the analysis of variance to determine whether a substantial relationship has been discovered. Calculate the pure error mean square and test for lack of fit of the model, and recalculate any variance ratios, if necessary F1,3,4,6
6. interpret the parameter correlations and parameter standard errors or t statistics to decide whether any variables may be deleted, F1,3,4,6
7. calculate multicollinearity statistics such as design or prediction efficiencies

and effective sample sizes to decide whether any variables may be deleted, F1,3,4,6

8. compare residual mean squares, F ratios, parameter correlations and variances, effective sample sizes, and the residuals to other fitted models to help decide whether to accept the model, F1,3,4,6
9. summarize the computer output and findings and decide either to accept the fitted model or to change it. Suggest how it should be changed, and why, F1,3,4,6
10. calculate marginal and joint confidence regions for parameters and test any hypotheses required. Calculate the effective sample sizes for the variables to indicate the extent of collinearity and explain these. F1,3,4,6
11. compute confidence limits for the response function and for future observations at selected or specified points, and provide expressions for the limits to enable simple calculation at other points F1,3,4,6
12. use the reserve data, or more recent data, to validate the model by comparing the actual values with predicted values and their confidence limits, or the new parameter estimates and confidence regions with the old, F1,3,4,6
13. interpret the model and its implications in terms of the experimental situation. If anything "doesn't make sense" or is "startling" initially examine this aspect carefully and justify it or reject the model. If there is more to be examined, but this can only be done by getting more data, describe the experiments which would be most informative.

VII Writing the report

Given the previous information, the student will be able to:

1. assemble and organize the computer output, plots and summary information and use this to develop an outline for the report,
2. prepare rough drafts of the report, describing the important parts of the analysis, the findings, including the final model plus parameter standard errors or confidence regions and any especially important points such as any unexpected predictor variables, any model or experimental design limitations, and suggestions for future experiments F1,3,4,5,6
3. calculate design and predictability efficiencies for the model and design for specified points and explain them, F1,3,4,5
4. modify and revise the drafts until a satisfactory report is obtained; the report should include important computer output and plots with analyses and derivations as appendices. Graphs and figures should be properly labelled and titled, and good referencing and sectioning of the report should be done. There should also be a clear, factual summary, and a list of references, if necessary.

NON-STATISTICAL ASPECTS OF STATISTICAL CONSULTING

Douglas A. Zahn and Daniel J. Isenberg
Harvard University

ABSTRACT

Statistical consulting is a complex activity that requires statistical and non-statistical skills on the part of the consultant. Both sets of skills, in conjunction, determine the ultimate success of the consultation. This paper focuses on the non-statistical aspects of the initial meeting between statistical consultant and client, presenting a model for dividing that session into four parts:

A. information gathering by the consultant,
B. defining the assistance sought by the client,
C. deciding on first actions to be taken by the client and the consultant,
D. discussion of the client's and consultant's expectations of the method (e.g., renumeration, joint publication).

Each part gives rise to specific issues and requires various nonstatistical skills on the part of the consultant to deal with these issues. We believe that the effectiveness of the consultation can be enhanced by learning and practicing this model and the constituent skills.

INTRODUCTION

Statistical consulting poses some difficult challenges, some of which have little to do with statistics per se. Occasionally, for example, one encounters a domineering client who makes unrealistic demands. This client typically charges through the door ten minutes late and informs the consultant that s/he wants all of their statistical work done by the consultant, including a substantial amount of programming, coding and keypunching, and "Would you please have it ready by this afternoon?". Not infrequently, it turns out that this client is a good friend of one's boss, sits on the advisory board to the academic community or the firm in which one is working.

Some students in a university or junior staff in industrial firms also present difficult consultant situations. Having had one introductory statistics course, they have decided to tackle a thesis or research topic that involves statistical analyses far beyond their budding expertise. Their advisor or supervisor, also possessing little statistical expertise, has informed the impatient student or junior staffer that their problem surely involves heteroscedasticity, oblique rotations, and principal components. Not wanting to appear ignorant, this person greets the consultant with "Hello. Do you think you have ten or fifteen minutes to help me with a multivariate problem that has multicollinearity and suppression?".

A third issue which arises in most, if not all, consulting situations is the negotiation (or lack thereof) of credit for work being done, amount, form and scheduling of payments, authorship, use of data sets, confidentiality, division of responsibilities and so on. If these questions are not addressed directly and skillfully, much time, effort, and goodwill can be wasted. However, both clients and consultants frequently answer these questions by default, either because they are too hot to handle, or because both parties assume (often falsely) that the answers are clear and agreed upon.

Evidence available from many sources suggests that at least some individuals who are recent recipients of a Ph.D. degree in statistics do not have the skills necessary to successfully resolve the above problems and others like them that arise in the course of statistical consulting. Deficiencies in non-statistical skills are not only found in novice consultants, but in some senior statisticians as well, according to our conversations with clients in industrial and university settings.

Sprent (1970) also questions whether statistical consulting is as effective as it should be. He suggests several possible reasons for deficiencies here, including "There is little or no training in the art of consultancy" (p. 140). Though many courses in statistical consulting have developed during the ten years since he wrote this, there is still widespread sentiment that "the amount of formal training that can be given in the 'art' of consulting is rather limited. This is more a matter of apprenticeship" (p. 142). Though other authors, including Cox (1968), also discuss the "art" of consulting, this term is generally left undefined, retaining a magical, mysterious air. Perhaps these authors mean to include in the "art" of consulting those things that one does in a consulting session that are not technical statistical things, but which do help to produce high quality work.

We also make a distinction in this area. We distinguish between the statistical work relating to the problem under consideration which is generally done outside the meeting between statistical consultant and client and the statistical consulting per se which occurs during the meeting. Thus, the statistician engaged in this type of activity definitely does statistical consulting and may do statistical work for the client.

The purpose of this paper is to contribute to the reduction of the mentioned deficiencies by proposing a model for effective statistical consulting. We suggest that the model can serve as a road map in identifying non-statistical issues which arise during consulting sessions and in specifying some of the skills required to deal with these issues. As such, we feel that it can serve to reduce the time required to begin to learn the "art" of statistical consulting. Though our experience is primarily, though not exclusively, in a university context, we believe that the model proposed in this paper

can be useful in statistical consulting in any context, industry, government, consulting firm, or university (and may be applicable to other types of expert consulting - legal, financial, architectural and so on).

One major underlying assumption of our approach to statistical consulting is that the most important activity statistical consultants (or helpers of many kinds) can engage in is to help clients understand the nature of their statistical and/or methodological problem in such a way that the clients themselves can take appropriate action and learn something in the process (sometimes called a "problem-minded" approach, as opposed to "solution-minded", Maier, 1970). Thus, statistical consulting from our perspective has two goals: (a) solving an immediate statistical or methodology-related problem, and (b) increasing the capacity of clients to understand and solve similar problems when they arise in the future. The consultant and client both may have needs to learn new things, master new subject matters, and develop self-confidence by tackling new issues. By emphasizing increasing client competence we hope to encourage both consultants and clients to pay attention to this aspect of consulting. Note that the immediate problem solved by the consultant and client may differ from the problem initially presented by the client (see Schein, 1969). We also note that there may be aspects of the environment in which one is consulting which cause one to set goals in addition to or in replacement of those just stated. Argyris (1970) suggests similar conditions for effective consulting, namely that a consultant (he uses the term "interventionist") must maintain or increase clients' current performance while increasing their capacity to generate valid information, free choice and internal commitment in solving problems in general. We place somewhat more emphasis on solving the immediate problem than does Argyris but are in general agreement with his formulation.

An early assessment of the role of the statistical consultant is given by Eisenhart (1947) who maintains that the consultant shall determine "what conclusions can be drawn and with what confidence" and "the amount and type of data needed as evidence". We concur that these aspects of statistical consulting are important and are embodied in our Goal (a) above. Our goals expand on Eisenhart's view by including an increase in the competency of the client as another desired outcome of statistical consulting.

As a final introductory comment, we emphasize our assumption that the consultant has the appropriate level of formal training in applied statistics, mathematical statistics, numerical analysis, mathematics, computers, etc., as is required to be competent in his or her job. We also emphasize, as do Cox (1968) and Daniel (1969), the importance of serious exposure to the scientific method and to how scientists actually work. We do not suggest that good consulting skills can make up for a lack of statistical competence or depth in these other areas; they cannot. Rather, we suggest that good consulting skills can amplify the utility of the statistical competence one already has.

A MODEL FOR STATISTICAL CONSULTING

We describe in this section the application of our model for effective consulting to the first session between consultant and client. One of our reasons for doing this is to meet the request of many supervised consulting students over the years for a description of what actually happens during a consulting session. The first session is particularly important because it establishes the behavioral and attitudinal norms for the rest of the consulting relationship. Furthermore, the skills that are useful in the first session and the model we propose are also valuable in later sessions. This session may be one of the most difficult between client and consultant because it brings together two people who have probably never met before and who may well have negative expectations of each other and of the outcome of the encounter (Hyams, 1971).

Partly borrowing from Fosmire and Wallen (1971) (see also Kepner and Tregoe, 1965), our proposed approach to the first session involves four parts:

(a) identifying relevant aspects of the problem situation;
(b) identifying and formulating the client's goals or targets;
(c) deciding upon the actions to be taken (proposals); and
(d) negotiating various aspects of the consulting relations (see Harrison, 1973; Sherwood and Glidewell, 1971).

Each of these four parts raises various issues and requires specific skills on the part of the consultant to resolve these issues successfully. The parts of the session are not necessarily executed in this order; also, iteration is certainly possible. To reiterate, we encourage an initial and prolonged emphasis on the situation and targets in order to arrive at effective proposals which are executed in accordance with the results of the negotiation. (See Maier (1970) for evidence that this is an effective strategy for problem-solving.)

Our observation is that, in the absence of this model, frequently consultants focus on statistical solutions to the client's problems, resulting in (a) the right answer to the wrong problem, or (b) the client going off to blindly perform analyses that s/he does not understand. Possibility (a), referred to as an "error of the third kind" by Kimball (1957), may be especially difficult to detect, especially for inexperienced consultants. To make matters worse, consultants frequently do not find out that either of these possibilities has occurred because some clients simply do not come back, or if they do come back, they do not directly inform the consultant about the bad advice they received. (See Argyris and Schon, 1974, for some reasons why this happens.)

Identifying the situation: The initial part of the first consulting session involves determining the current state-of-affairs with respect to the client's specific problem and the client's overall situation. In addition to understanding whether the client has collected three or four waves of data over ten years, it is also important to understand if the client is under time-pressure, if there are other collaborators, if the client has complete autonomy over the use of the data, if similar research has been conducted elsewhere, and so on. Hyams (1971) notes that there is a benefit in beginning the consultation with a sincere expenditure of effort on the consultant's part to learn the substantive difficulties in and importance of the work being done. This may contribute to allaying the client's fears or negative expectations of statistics and statisticians.

Both clients and consultants have initial assumptions about what information is relevant and irrelevant; often clients decide to present only limited aspects of their problem. Other clients may do quite the opposite, innundating the consultant with all possible details. A direct approach is for the consultant to state at the outset what s/he needs to hear, the range of issues that are relevant, and then ask the client if there is anything that they think is particularly relevant that the consultant has not mentioned. Cox (1968, p. 795) presents and discusses an especially helpful list of statistical topics to examine in this part of the consulting session,

a. objective of the experiment, hypothesis formulation and specification
b. experimental plan and design
c. execution
d. analysis
e. interpretation and inference.

As the client defines the situation, the consultant should frequently reflect back to the client his or her own emerging understanding of the client's situation. Following is an interchange in which the statistical consultant (SC) reflects back information the client (CT) has presented so as to clarify the situation:

CT: The 75 subjects formed three organizations and played the management game for seven sessions.
SC: So, for session one you divided the subjects into three groups and ...
CT: No, that was done randomly.
SC: You randomly divided the subjects into three groups and each group played the management game.
CT: Yes.
SC: And then for session two it was the same thing.
CT: Yes.
SC: You again randomly divided the subjects into three groups and they played the game again.
CT: No. The same groups were used in the second session.
SC: I see.

Note the clarification that occured at various points as a result of the SC "merely" reflecting the comments of the CT. Such behaviors, called "active listening", "reflective listening", or "paraphrasing" (Egan, 1975; Jung, Howard, Emory and Pino, 1972), help prevent the misunderstanding and ambiguities that are so common in any conversation. Reflective listening not only helps the consultant get a clearer understanding of the client's problem situation (and also provides a chance to catch one's breath!), but it also helps the client check to see if his or her intended communications are getting across.

Another tool which may be of use to the consultant is the probing question (e.g., "Did you randomize the assignment of subjects to treatments?"). This may be of particular use after the client has had a chance to define the situation as s/he sees it (see Kahn and Cannell, 1957).

Some clients start their description of their problem with a solution. "I just have one question that I want you to answer for me. Shouldn't I perform MANOVA on this data set?". Occasionally, it is a legitimate consulting function to essentially endorse clients' solutions to their own problems; such initiative and independence on the part of the client should be encouraged by the consultant. However, clearly the consultant's most pressing job in this situation is to first understand the problem, grasp what questions or goals the client has in his or her analysis, and only then to give the endorsement. Consultants, in particular beginners who may feel pressure to provide crisp, immediate answers, will enhance their survival probabilities if they learn to say to clients who poke their heads in the door "On the surface that sounds plausible, but I don't have enough information about what you are trying to do in order to give you a thoughtful answer. Why don't we set up an appointment to explore that?". This is not always what the client is looking for, but the alternative is for the consultant to give a half-baked answer and over the long-run to loose credibility among clients.

Defining Targets. At some point both client and consultant will be satisfied that they have enough understanding of the situation to proceed to target definition. If the situation has been well described, defining appropriate and feasible goals will proceed smoothly, the first thing to check is that the situation is clearly and thoroughly described.

There are several issues that can emerge when client and consultant set out to clarify, formulate, and (perhaps) reformulate the client's research questions. Perhaps the most difficult issue to face is when the client's data do not appropriately address the research objectives. Related to this are the more common instances when a client's theory is not clearly enough thought through in order to define appropriate research objectives. Whereas the only ultimate solution to these problems is to provide better methodological education throughout the organization, the skillful consultant can offer some

short range assistance. Between the Scylla of passive acceptance ("Oh, so you want to find out if A causes B by correlating them?") and the Charybdis of disdainful punishment ("Why, didn't you ever learn in Statistics 100 that you can't do that?!") is a response of constructive confrontation that informs the client that there is a gap between the desired goals and the resources available. The consultant can help the client explore how the critical mistake was made, how to redesign the study, or which research questions are addressable with the existing data set.

We believe that consultants can play an invaluable role in "simply" helping clients to formulate unclear research objectives. Questions such as, "What would you hope to answer by performing that analysis?", or "How will you know if you've answered that question?" (see Isenberg and Fosmire, 1977) lead the client to define goals that are much more in accord with available theory and resources.

Formulating Proposals. As with defining targets, clients and consultants who experience difficulties specifying actions for the client to take should realize that these difficulties are frequently due to unclear situations and targets. That is, the clearer the problem is, the clearer and more creative (Maier, 1970) the solution will be. The above not withstanding, difficulties do arise in formulating proposals per se, the most frequent being that the most appropriate statistical analyses sometimes require knowledge and ability that the client does not have. Two solutions are: (a) for the consultant to perform a state-of-the-art analysis for the client, and (b) for the client to become better educated in statistics. Both solutions have strengths and weaknesses. The first solution may be most painless but does not increase the capacity of the client to solve future statistical problems. The second solution increases the client's capacity but does not solve the client's immediate need.

We advocate a third solution: the consultant helps the client formulate statistical analyses which are within the grasp of the client, while fully informing the client that the consultant is doing this and why. Consultants can feel perfectly free explaining the rationale behind this strategy. "In your case, client X, there are three possibilities ... [consultant then outlines the three solutions]. I will concentrate on analyses that you yourself can understand and perform, because that way you yourself will be learning something and you yourself will be able to interpret and define the results. To the extent that there are faulty assumptions underlying your analyses, I will tell you about them and will try to give you some notion as to how much these might affect your analyses". This solution contributes to achieving one of our major goals in statistical consultation, namely increasing the client's ability to deal with this class of problems. Education of the clients is also mentioned as an important goal by Hyams (1971). He notes that this is another way of overcoming a client's defensive rejection of the application of statistics to the work being done.

Of course, in some situations it is necessary to do a state-of-the-art analysis that is beyond the client's abilities. Using our model the client and consultant would discover this while discussing actions to be taken. This situation poses a dilemma with respect to our two goals for statistical consultation in that it essentially forces there to be a collaborative relationship between the statistical consultant and the client in which the consultant engages in statistical work in addition to statistical consulting. If the statistical consultant is doing all the statistical work, a natural question is "How are the competencies of the client being increased?". Even in collaboration, the client may still learn some statistical material and thereby achieve Goal (b); the question is how much and at what level does the client wish to learn.

At this stage clients need to be able to think their actions through, and consultants can be helpful by simply sitting silently while the client thinks quitely or out loud. Whereas this may seem overly simplistic, adults in our culture often feel uncomfortable when silence befalls a conversation. We wish to legitimize silent reflection as a useful strategy for joint problem-solving (see Van de Van and Delbecq, 1975). Another technique is to have clients run through their next steps out loud to the consultant. Not only do consultants learn how much of their proposal is comprehended by the client, but also clients get a chance to practice their own reflective listening skills!

Statistical analyses provoke anxieties in many clients, partly because most people avoid trying to appear very ignorant, and partly because there is so much new to learn in order not to be. This puts clients in a difficult bind - damned if they show their ignorance and damned if they don't. Consultants must try to create the conditions where the answer to "Is all that clear?" is not "yes" when it should not be. (Consultants who feel pressure to know the answers to all problems may make it more difficult for clients to identify their own ignorance. Not only clients, but also statistical consultants can say "I don't know, I'll look that up." However, consultants who admit ignorance as a ploy help to create the opposite dilemma - clients who are afraid to admit areas of competence.)

Discussing Expectations. The most difficult part of the first session for both consultants and clients often is the discussion of some topics that are generally regarded as, if not taboo, at least very sensitive - money, credit, mutual expectations, division of labor, and so on. Clients, for example, may expect the statistical consultant to be research assistant, statistical advisor, collaborator, keypunch operator, or computer programmer. On the other hand, consultants may expect clients to commit themselves to a minimum number of consulting sessions, to read some methodological articles, or allow the consultant to use the data set for research. Expectations exist; the consequence

of not discussing them openly may be a belated discovery that client and consultant expectations are quite diffierent. One of our colleagues found himself six months into a consulting project when it became clear to him that a footnote at best would be his desserts, whereas he had all along expected co-authorship. The discussion or negotiation being done in this part of the consulting session may range from informal, resulting, e.g., in a verbal agreement that there will be another meeting soon, to formal, resulting in a written legal contract.

When consultants and clients openly discuss expectations, conflicting expectations may become apparent. Consultants do not have to always have their way, nor is "the client always right". With patience and careful, reflective listening, client and consultant can often identify a modus operandi which is substantially within the expectations of both. For example, instead of doing the programming, the consultant may be able to locate someone who will. Or, if the client's budget is limited, perhaps it can be re-allocated to cover the most crucial statistical analyses (see Filley, 1975, for an approach to resolving these types of conflicts).

Particularly salient in discussing expectations are power and status differences between the consultant and the client. The consultant may be the client's statistics instructor, or the client may be a prestigious professor at the university where the consultant is a graduate student. If we are to believe sociologists and social psychologists, no amount of open discussion will completely eliminate the effects of status differences. We believe that it is essential for consultants to have a very clear picture of their organizational position, career aspirations, job responsibilities, alternative employment possibilities, as well as an explainable rationale for dealing with clients in a way that will achieve both goals of statistical consulting - short-run problem-solving and long-run capacity of clients to solve their own problems. Obviously, these are issues to be thought out before the consulting session. Sessions with one's colleagues and supervisors, as well as with oneself, are required to develop positions on these issues.

Hyams (1971) points out another reason for considering these issues in advance. He points out that consultants desire "the three R's: Reality, Respect, and Reward" (p. 202) and emphasizes that the consultant must constantly be evaluating his or her own motivation and expectations, in addition to those of the client. He also notes the special perils which await the young consultant in this regard. Often this consultant receives the most difficult clients, which makes it all the more critical to formulate a position with respect to the three R's and the handling of difficult clients.

Deming (1965) presents a comprehensive description of his position on key matters to be dealt with by consultant and client when they are discussing expectations. Dealing with all the mattters Deming mentions may be difficult in the short run, but will almost surely pay off in the long-run.

EASIER SAID THAN DONE

Our approach is much easier to write about than it is to practice, especially when consultants' and clients' concerns about themselves and their competence begin to enter into the relationship, as they inevitably do (see Argyris, 1976). Specifying that in order to conduct effective consulting sessions one needs to directly discuss expectations is a bit like specifying that in order to ride a bicycle one must develop the skill of maintaining one's balance. Such a prescription is mainly helpful in retrospectively explaining to the experienced bicyclist what s/he did correctly; it is of questionable use for teaching the novice. However, we hope that our model will be useful (as it is to us) in that it directs attention to some of the more relevant behavioral variables in teaching and practicing effective statistical consulting. In learning how to ride a bicycle, there is no manual which by itself can teach such a skill. Similarly, in consulting there is no substitute for practice, quality experience, observation of experts, and competent supervision.

REFERENCES

Argyris, C. Intervention Theory and Method: A Behavioral Science View. Reading, MA: Addison-Wesley, 1970.

Argyris, C. Theories of action that inhibit individual learning. The American Psychologist, 1976, 31(0), 638-654.

Argyris, C. and Schon, D. A. Theory in Practice: Increasing Professional Effectiveness. San Francisco: Jossey-Bass, 1974.

Bliss, D. I. Communication between biologists and statisticians, a case study. The American Statistician, 1969, 23, 15-20.

Boen, J. R. The teaching of personal interaction in statistical consulting. The American Statistician, 1972, 26, 30-31.

Boen J. R. and Fryd, D. Six-state transactional analysis in statistical consultation. The American Statistician, 1978, 32, 58-60.

Boen, J. R. and Smith, H. Should statisticians be certified? The American Statistician, 1975, 29, 113-114.

Bross, I. D. J. The role of the statistician: Scientist or shoe clerk. The American Statistician, 1974, 23, 126-127.

Brown, T. H. The statistician and his conscience, with discussion by W. W. K. Freeman. The American Statistician, 1952, 6, 14-20.

Cameron, J. M. The statistical consultant in a scientific laboratory. Technometrics, 1969, 11, 247-254.

Court, A. H. Statistical standards, with discussion by J. E. Morton. The American Statistician, 1952, 6, 6-7.

Cox, D. F. Some observations on the teaching of statistical consulting. Biometrics, 1968, 24, 789-802.

Daniel, C. Some general remarks on consultancy in statistics. Technometrics, 1969, 11, 241-246.

Delbecq, A. and Van de Ven, A. H. Group Techniques for Program Planning: A Guide to nominal Group and Delphi Processes. Glenview, IL: Scott-Foresman, 1975.

Deming, W. E. Principles of professional statistical practice. Annals of Mathematical Statistics, 1965, 36, 1883-1900.

Egan, G. The Skilled Helper: A Model for Skilled Helping and Interpersonal Relating. Brooks-Cole, 1975.

Eisenhart, C. The role of a statistical consultant in a research organization. Proceedings of the International Statistics Conference, 1947, 3, 309-313.

Filley, A. Interpersonal Conflict Resolution. Glenview, IL: Scott-Foresman, 1975.

Fosmire, F. R. and Wallen, J. The STP model of problem-analysis and decision-making. Mimeo, University of Oregon, 1971.

Freeman, W. W. K. Training of statisticians in diplomacy to maintain their integrity. The American Statistician, 1963, 17, 16-20.

Gibbons, J. D. A question of ethics. The American Statistician, 1973, 27, 72-76.

Haley, J. and Hoffman, L. (eds.) Techniques of family Therapy. New York: Basic Book, 1967.

Hansen, M. H. Statistical standards and the census, with discussion by C. D. Long. The American Statistician, 1952, 6, 7-14.

Harrison, R. Role negotiation: A tough-minded approach to team development. In W. Bennis, D. Berlew, E. Schein, and F. Steele (eds.). Interpersonal Dynamics (3rd edition). Homewood, IL: The Dorsey Press, 1973, 467-479.

Hyams, L. The practical psychology of biostatistical consultation. Biometrics, 1971, 27, 201-211.

Isenberg, D. J. and Fosmire, F. R. Utilizing general and specific information in joint inquiry. Unpublished manuscript, Harvard University, 1977. (Available from senior author.)

Jung, C., Howard, R., Emory, R., and Pino, R. Trainer's Manual: Interpersonal Communication. Portland, OR: Northwest Regional Educational Laboratory, 1972.

Kahn, R. L. and Cannell, C. F. The Dynamics of Interviewing: Theory, Technique, and Cases. New York: John Wiley & Sons, 1957.

Kepner, C. H. and Trego, B. B. The Rational Manager: A Systematic Approach to Problem-Solving and Decision-Making. New York: McGraw-Hill, 1965.

Kimball, A. W. Errors of the third kind in statistical consulting. Journal of the American Statistical Association, 1957, 57, 133-142.

Maier, N. R. F. Problem-Solving and Creativity in Individuals and Groups. Brooks-Cole, 1970.

Salsburg, D. S. Sufficiency and the waste of information. The American Statistician, 1973, 27, 152-154.

Satir, V. Conjoint Family Therapy: A Guide to Theory and Technique (rev. ed.). Palo Alto, CA: Science and Behavior Books, 1967.

Schein, E. Process Consultation: Its Role in Organization Development. Reading, MA: Addison-Wesley, 1969.

Sherwood, J. J. and Glidewell, J. C. Planned renegotiation: A norm-setting OD intervention. In W. Bennis, D. Berlew, E. Schein, and F. Steele (eds.). Interpersonal Dynamics (3rd edition). Homewood, IL: The Dorsey Press, 1973, 480-487.

Sprent, P. Some problems of statistical consultancy (with discussion). Journal of the Royal Statistical Society, Series A, 1970, 133, 139-165.

STATISTICAL TEST ITEM COLLECTION SYSTEM

A.D. Bugbee, H.D. Merrill, J.A. Warren, University of New Hampshire

INTRODUCTION

During 1973 a group of faculty members at the University of New Hampshire formed an ad hoc committee to study the content and consistency of the many introductory statistics courses offered on campus. They proposed that a single intercollege course, utilizing a modular format, computerized testing, and video-taped lectures, would have distinct advantages over the existing situation. The efforts of these faculty resulted in the undertaking of two related projects. The first project was the development of such an intercollege course, using instructors drawn from a variety of departments, a large tutor staff, and competency-based testing. (For further information, see Paper #5 listed in the appendix.) The second project, entitled "Diagnostic and Instructional Support Services for College Students of Statistics", was initially funded by the Teaching and Learning Council at the University of New Hampshire, and later was funded by the National Science Foundation.* This paper will report on a major element of the latter project, "The Statistical Test Item Collection System", or STICS.

STICS was developed as a support system for instructors to assist in both the specification of introductory statistics courses and the delivery of exams by sampling item collections. The system consists of two main components:

1) a data base of over 3000 introductory-level questions and answers; and

2) a collection of conversational computer programs.

The intent of this paper is to explain the make-up of each of the above components, and to illustrate how each can be used to assist an instructor of introductory statistics. It should be pointed out that both components are still under development, and suggestions, comments, or criticisms are welcome.

THE DATA BASE OF QUESTIONS AND ANSWERS

The data base currently contains 3000 questions with answer guidelines for introductory statistics courses. These questions have been obtained from a wide variety of sources, including approximately fifty individual professionals. The first solicitation of contributions for the data base consisted of a large mailing conducted in June, 1974. It was sent to departments offering introductory statistics courses at institutions throughout the United States and Canada, and invited instructors to send in copies of final exams or proficiency tests.

*National Science Foundation Grant SED A01 76-12191.

A major part of the current data base can be attributed to the numerous responses to this mailing. Since that time, items have been submitted by various individuals, some on our urging and others on their own initiative. Another source of items has been the question set assembled to meet the objectives of the previously noted modular introductory statistics course developed at the University of New Hampshire.

In retrospect, several advantages and disadvantages of this method of collection have become apparent. One advantage was that inspection of the format and content of questions received made it possible to provide topic coverage reflecting a broad range of viewpoints. It was hoped that this diversity of sources would assure "something for everyone" in terms of providing items useful to instructors of courses designed for varying purposes and audiences. Unfortunately, this process also contributed to the accumulation of items of inferior quality and uneven topic coverage, such that some topics were addressed repeatedly while others were totally neglected. During the final year of this project, the current collection will be edited and corrected, and an additional 1000-1500 items will be added. New items will be selected so that topic coverage will be more evenly distributed.

The following procedure has been established for screening and preparing items for inclusion in the data base. Each question is first judged as to whether it is truly "introductory" in nature. Appropriate items are then checked for answer guidelines. If an answer has been supplied by the contributor, it is verified for accuracy and completeness. Otherwise, an answer guideline is written. It is hoped that these guidelines not only indicate the correct answer, but can also provide a learning aid to any student who has incorrectly answered the question.

The next step in the process requires that the question be prepared for entry in the data base. Preparation entails classifying each item according to various keyterms and question characteristics. The keyterms describe the topical content of the question, i.e. regression, sampling, etc. The question characteristics provide information on the following:

1) cognitive level,
2) math prerequisite,
3) area of application,
4) difficulty level,
5) time required,
6) question type,
7) equipment needed (calculator, computer, none),
8) whether a chart or graph is needed that is not included,
9) whether the question involves multiple parts, and
10) whether a statistical table is needed.

The keyterms, characteristics, question, and answer guideline are then entered and stored in the computerized data base. Following this, each question is subjected to a review process that deliberately includes several different viewpoints. Each question is reviewed by: the project editor, a graduate student in statistics, a member of the American Statistical Association, and a person who is not a member of the ASA. The intent of the review process is to eliminate poor wording in questions, incorrect answers, and/or typing mistakes, prior to making the questions available for general use.

One complication that arises because the computer is being used as the storage device is the use of mathematical symbols and notation. Such characters as summation symbols, integral symbols, Greek letters, superscripts, and subscripts have not been available on most computers. This problem has necessitated the use of some special notational conventions. Although at first inspection some of these conventions seem awkward and hard to read, our experience has shown that most students can adapt to them without much difficulty. However, we are still not pleased with this situation and are currently seeking alternatives.

SYSTEM OF PROGRAMS FOR COMPUTER-ASSISTED TEST CONSTRUCTION

The system designed for computer-assisted test construction consists of a series of self-documented, conversational programs. Although these programs were originally developed for interactive use, there now are versions in use on batch systems. These programs were designed for use with the data base of statistical questions. However, they also may be used for item collections dealing with other content areas (assuming that compatible data structures are used).

STICS provides three main types of programs for managing an item collection:

1) A program (SELECT) for extracting items from the data base. The purpose of this program is to allow a user to designate keyterms and question characteristics to specify questions for use in a complete course, a module, a unit, a test, or simply a homework assignment. This program retrieves the questions that satisfy these requirements, and forms a file of this subset of the data base. The subset can be printed out and used as a reference, or kept on-line and used later to generate exams. Files resulting from the use of SELECT are usually considerably smaller in size than the full data base, and consequently much easier to work with. A single course might make use of several subsets corresponding to different portions of the course. The process of forming such sets is discussed further in papers by Warren and Michener.

2) A set of editing programs. These programs allow a user to refine a subset of questions. They provide for adding, removing, or modifying questions, answers, keyterms, and question characteristics. In general, the editing programs allow a user to "improve" on what has been selected from the complete data base.

3) An exam generating program (BEGET). This program provides for the use of random sampling to prepare exams and answer guidelines. If the exam items to be used for a course or a part of a course have been identified and put in a subset, BEGET can be utilized to generate exams by sampling a file for that subset. Keyterms or question characteristics can be used when choosing items to appear on an exam.

AVAILABILITY OF PROGRAMS AND DATA BASE

These programs have been written in ANSI Standard Fortran, and have been run on a DECsystem-10 at the University of New Hampshire. Other versions are available for a DECsystem-20 and for IBM 360/370/3031 systems. All versions are distributed, with the data base, via magnetic tape. Costs involved are minimal - copying charges, postage, etc.

The data base is available in machine-readable form that can be used with computer programs for retrieval and printing, or as a catalog printed on microfiches. The catalog includes several indices that make it possible to search for questions according to keyterms or question characteristics.

WE NEED ASSISTANCE

We need assistance, in one or more of three activities, to improve the data base.

1) We need individuals who are willing to act as reviewers. This involves proofing questions and answers for accuracy.

2) We need contributors who will submit 5- 20 items that they feel would be useful to other statistics instructors. Our goal is to fill gaps in the present item collection.

3) We need individuals who are willing to scan the microfiche and point out the inferior items or content areas that need attention. We would consider it a bonus if they were willing to report on actual classroom use of individual items.

In short, we need feedback. Suggestions, comments, and criticisms are welcome.

SUMMARY

STICS consists of a data base of introductory statistics questions and computer programs for test construction. Our aim in developing this

system was to provide aid for the process of forming question frames and producing exams for introductory statistics courses. These aids are intended to reduce the start-up time and effort required to identify question sets that agree with the educational objectives of a statistics course. Such question sets then provide a basis for using sampling techniques to prepare exams. STICS provides aids for dealing with some of the troublesome problems that make it difficult to achieve competency-based testing, or courses that are consistent in what they require of students.

For further information on any of the above, contact:

Alvin D. Bugbee
Office of Biometrics
304 McConnell Hall
University of New Hampshire
Durham, NH 03824

Telephone: (603) 862-1990

APPENDIX

PROJECT PUBLICATIONS AND PRESENTATIONS*

Papers

1. "Course Objectives and Planned Comparisons" by Jerry A. Warren. Paper presented at the joint meetings of the American Statistical Association, the Institute of Mathematical Statistics, and the Biometric Society, San Diego, California, August 14-17, 1978.

2. "Automating Exams for a Statistics Course I - The Beginning of a Case Study" by J.A. Warren, R.D. Michener, and A.D. Bugbee. Paper presented at the 16th annual convention of the Association for Educational Data Systems, Atlanta, Georgia, May 16-18, 1978.

3. "Automating Exams for a Statistics Course II - A Case Study" by R.D. Michener, A.D. Bugbee, and J.A. Warren. Paper presented at the 16th annual convention of the Association for Educational Data Systems, Atlanta, Georgia, May 16-18, 1978.

4. "A System for Serving Faculty and Students of Introductory Statistics" by William E. Geeslin. Paper presented at the NAUCAL meeting, November, 1977.

5. "The Computer as an Aid for Modular Instruction in Statistics" by William E. Geeslin. Paper presented at the eighth conference on Computers in the Undergraduate Curricula, Michigan State University, June, 1977.

6. "Diagnostic and Instructional Support Systems for Students Taking Introductory Statistics" by George Miaoulis. Paper presented at the annual meeting of the American Mathematical Association of Two-Year Colleges, Atlanta, Georgia, October 13, 1977.

Technical Reports

7. "Data Bank Procedures for the Diagnostic System, Technical Report No. 1" by W. Geeslin and A. Bugbee. Unpublished report, October, 1977.

8. "Resource System, Technical Report No. 1" by G. Miaoulis and D. Baird. Unpublished report, October, 1977.

Manuals

9. SELECT User Manual by H.D. Merrill.

10. BEGET User Manual by H.D. Merrill.

11. STICS Editing Programs User Manual by H.D. Merrill.

12. Statistical Test Item Collection System Programs for Batch Environments by H.D. Merrill.

13. System Operator's Manual by H.D. Merrill.

14. Supplementary Documentation by H.D. Merrill.

Miscellaneous

15. "Update - A Newsletter About the University of New Hampshire Statistics Teaching Project" - May, 78, No. 1; October, 78, No. 2; April, 79, No. 3.

16. Microfiche copy of questions and answers for Introductory Statistics. (A fee of $5.00 charged to people not contributing or reviewing questions.)

*Copies of the materials listed above can be obtained from: The Office of Biometrics, 304 McConnell Hall, University of New Hampshire, Durham, NH 03824.

ON SELECTING BEST PREDICTOR QUESTIONS IN A STANDARDIZED PASS-FAIL EXAMINATION

Jack E. Graham, Lawrence E. May, Habeeb A. Sheikh, Carleton University

1. Introduction

It was recognized by a number of Carleton University instructors lecturing in first year calculus and linear algebra courses that the level of comprehension achieved by many of their students was less than satisfactory. They believed that the difficulty in teaching Freshman-level mathematics lay not with the students' previous exposures to these subjects in high school or elsewhere but rather with their command of basic mathematics.

In 1974 a 20 question multiple choice quiz was developed that covered material which is usually encountered in all years of pre-university mathematics study; major emphasis in the test is placed upon pre-calculus material considered to be useful (and often essential) to the development of abilities in and understanding of calculus. Thus, much of the test is concerned with the manipulation of functions, basic algebra, exponents, trigonometry and logarithms. The 50 minute test is given prior to and during the university registration period. All students deemed to have "passed", i.e., obtained a score of 9 or more out of 20, are permitted to enroll in the mathematics course of their choice while those "failing" are directed instead to a remedial level course for which full university credit is given. The "Placement Test" serves a second purpose as well; those students passing are further stratified into sections according to their demonstrated ability in mathematics with the "advanced stream" receiving an enriched course of study.

Needless to say, the introduction of a competency test was a controversial move with a number of the Departments affected claiming that either high school grades were certainly sufficient evidence or that the Mathematics Department was making excuses for its own past failures. An intensive examination of the September, 1976 test was undertaken with bibliographic information on students, test results, and subsequent performance in first year mathematics courses being collated. In this paper we give a summary of the major test results and then describe our attempts to use two different discrimination techniques to detect what subset of the 20 questions could be employed with reasonable reliability to distinguish those students with a satisfactory degree of basic mathematical skills. A fuller account of the Placement Test may be found in May, Graham and Sheikh (1979).

2. Test Results

The cumulative frequency and percent cumulative relative frequency distributions of the test scores for the 1170 students on which this paper is based are presented in Table 1. Note that 743 out of 1170 or 63.5% passed the quiz.

3. Success in Subsequent Mathematics Courses

Table 2 gives a cross-classification of the Placement Test scores by the pass rates in a subsequent first year calculus course intended for students not planning to major in mathematics. One set is computed on the basis of the number who actually wrote the final examination and the other also includes students who were absent from the final examination or withdrew from the course after the final "drop date". Some students scoring a 7 or 8 on the Placement Test were still permitted to avoid the remedial course if they were able to present a strong valid case to a Mathematics Department advisor. Conversely, some passing the Placement Test still preferred to take the remedial course. A strong relationship between the pass rates and the test scores is evident.

On the other hand, Table 3, which reports on the corresponding results for students taking the remedial course exhibits no such correlation, indicating that these students are more homogeneous in their basic mathematical skills (or lack thereof).

4. Discriminant Analysis of Placement Test Scores

Discriminant analysis addresses itself to the problem of assigning an individual to one of several possible populations of potential origin based upon a set of observations made on that individual. Fisher (1936) originally solved the problem for the two population case using an intuitive principle. However, his linear discriminant function assignment rule possesses optimum properties in terms of misclassification probabilities only if the vector of observed variables is assumed to be multivariate normal. In our application the vector of scores $(x_1, x_2, \ldots, x_{20})$ on the 20 Placement Test questions (where $x_i = 1$ if the i-th question is correctly answered and $x_i = 0$ if it is answered incorrectly or left blank) serves as the set of variables upon which assignment to either the "Pass" or "Fail" populations is based. Here the linear discriminant function is known to be $X = \Sigma_{i=1}^{20} x_i$ with classification rule "assign to "Pass" if $X > 8$ and to "Fail" if $X \leq 8$.

Linear discriminant analysis was used to study the interesting question "can a subset of the 20 Placement Test questions be used which will still exhibit good discriminatory characteristics?" Thus we are asking if all 20 questions are actually operative and seek a linear combination $Y = \Sigma_{j=1}^{q} a_j y_j$ (where y_j = some x_i, $i = 1, \ldots, 20$, $q < 20$ and the a_j's are discriminant function coefficients) which approximates the X assignments well. The motivation for so-doing is three-fold: i) to eliminate questions that do not help to distinguish between the two groups; ii) to reduce the number of questions in the quiz, thereby decreasing the time required to complete it; iii) to permit (if required) the substitution of more sensitive questions for the ineffective ones. A stepwise selection procedure was implemented to achieve a ranking for combinations of one, two, three, ..., test questions in terms of their relative abilities to predict Pass and Fail. The algorithm commences by choosing the single variable (question) that has the maximum

discriminatory ability. Subsequent variables are chosen on the basis of their enhancement of the discriminatory power (i.e., reduce the "error sum of squares" to the greatest degree possible).

Since in this study 743 of the 1170 students writing the Test passed it, empirical prior probabilities of the two groups were taken to be 743/1170 = 0.635 and 427/1170 = 0.365. Table 4 cites the question entered at each stage of the selection process. (Space does not permit us to reproduce the Placement Test but the table still reports the question numbers as they appear in May, Graham and Sheikh (1979)). The subject matter area of each question, the percentage of students correctly answering the question and the "simplicity rank" (e.g., rank 1 = easiest question in terms of proportion of students correctly answering it) are also cited.

The "best" ten variable linear discriminant function was $Z = -2.16 + 0.463x_3 + 0.499x_5 + 0.421x_6 + 0.481x_8 + 0.320x_9 + 0.315x_{12} + 0.312x_{16} + 0.470x_{17} + 0.296x_{18} + 0.355x_{20}$, (1)

which can be approximately written as

$Z = -2.2 + 0.5(x_3 + x_5 + x_6 + x_8 + x_{17}) + 0.3(x_9 + x_{12} + x_{16} + x_{18} + x_{20})$. (2)

The classification rule is: assign to the "Pass" population if $Z \geq -0.42$ and to the "Fail" population if $Z < -0.42$.

Each of the 1170 cases was subjected to (1) in order to assess its discriminatory ability. Of the 743 that actually passed, 674 or 90.7% were correctly classified whereas of the 427 who failed, 367 or 85.9% were correctly assigned, yielding an observed misclassification probability of 0.110. This compares favourably with an 18 variable discriminator which correctly classified 94.1% of the "Pass" group and 92.0% of the "Fail" group for an overall 0.068 error rate. Because of the statistical argument used, the "best" 20 question discriminator will not be the simple aggregate $Y = \sum_{i=1}^{20} x_i$ (although it is well approximated) since in the discrete case no optimality properties are claimed.

What does it all mean? i) From Table 4 we observe that the three most difficult questions (14, 15, 19) are the three "least important" variables in terms of their ability to screen out those who should fail the Placement Test and are thus relatively inoperative on top of the others. ii) Similarly question 7, answered correctly by 92.2%, was entered at stage 17 and could be dropped as well. iii) No particular area covered by the Test appears to be of greater importance in discrimination. iv) Insofar as reverting to a 10 question Test, the misclassification probabilities will be increased but one might argue that the 20 question format incorrectly classifies some students anyway so the comparison is by no means clear cut. v) The approximate equation (2) of the ten available discriminator suggests that questions 3, 5, 6, 8 and 17 are of equal importance (with weights 0.5) and that question 9, 12, 16, 18 and 20 are also of equal importance (with weights 0.3), the former being somewhat more operative than the latter. This tends to be confirmed by the order of entry of the variables into the discriminant function with the second set following the entry of the first set. vi) Because Placement Test results are also used for streaming purposes, an inoperative difficult question might still form a basis for segregating "advanced stream" students.

Discriminant analysis techniques were employed to determine if any linear combination of variables could distinguish with high reliability those students who, having passed the Placement Test, would successfully complete the calculus course. Such variables as the Placement Test mark, year of the student in university, high school performance in two final year mathematics courses (calculus, functions and relations), and the number of years since the last mathematics course was taken by the student were used. Prior probabilities of pass and fail were set at 0.78 and 0.22, the ultimately realized rates. The only variables entered by the stepwise algorithm were the two high school marks and the Placement Test score. The resultant predictive capability was unsatisfactory; an empirical error rate of only 0.025 for the passes was realized but the error rate of 0.867 for the fail group is intolerable. These results agreed with regression analysis studies which attempted, with little success, to relate the final calculus marks to a similar set of background variables.

5. Nonparametric Question-by-Question Discrimination

An alternative nonparametric approach to identifying those questions that were most effective in distinguishing the passes from the fails was devised. Twenty 2-way tables with cell entries being the proportions passing and failing the Placement Test and also responding correctly and incorrectly to each of the questions were prepared. Taking guessing into account in the 5 alternative multiple choice format, one might regard the table

		Fail	Pass
Question Response:	Incorrect	0.8	0.0
	Correct	0.2	1.0

as indicating maximum discrimination by the question. With 1170 students writing the quiz and 743 passing, the "perfect" 2-way table would then give rise to a corrected chi-square value of 837.0. Such chi-square values were computed for each of the 2 x 2 tables and the values then used as indices to reflect the relative effectiveness of each question as a discriminator. As indicated in Table 5, question 5 with a chi-square value of 248.3 is assigned an effectiveness rating of 1, i.e., among all 20 questions it best discriminates the passes from the fails. For ease of comparison, the order of insertion of variables into linear discriminant function (from Table 4) is also cited.

The overall agreement under the two techniques is excellent. Apart from one question, the two ratings of the ten most effective questions agree and are more or less in the same order. (Comment: we realize that one should not attempt to place undue emphasis on insertion orders with a stepwise fit but nevertheless the agreement is striking since the algorithms have completely different logical bases.) Thus we can say with some confidence that questions 3, 5, 6, 8, 9, 12, 17, 18 and 20 would likely form

a sufficient subset to judge whether or not a student should directly proceed to study first year University calculus. Hence the Placement Test could be reduced in size and duration without too much loss in information.

REFERENCES

1. Fisher, Ronald A (1936), "The use of multiple measurements in taxonomic problems", Annals of Eugenics, 7, 179-188.
2. May, Lawrence E., Graham, Jack E., and Sheikh, Habeeb (1979), "The first year placement test at Carleton University, an investigation into its use and effectiveness", Carleton Mathematics Series, No. 157.

TABLE 1
PLACEMENT TEST DISTRIBUTIONS

Test Score	Cum. Frequ.	% Cum. Rel.Frequ.	Test Score	Cum. Frequ.	% Cum Rel.Frequ.	Test Score	Cum. Frequ.	% Cum. Rel.Frequ.
0	10	0.9	7	316	27.0	14	1026	87.7
1	13	1.1	8	427	36.5	15	1073	91.7
2	25	2.1	9	558	47.7	16	1100	94.0
3	51	4.4	10	667	57.0	17	1131	96.7
4	90	7.7	11	779	66.6	18	1154	98.6
5	124	10.6	12	863	73.8	19	1167	99.7
6	197	16.8	13	959	82.0	20	1170	100.0

TABLE 2
FRESHMAN CALCULUS PASS RATES VS. PLACEMENT TEST SCORES

Test Score	(No.Pass)/(No.Writing)	%	(No.Pass)/(Total No.)	%
7	6/15	40.0	6/19	31.6
8	11/30	36.7	11/35	31.4
9	46/78	59.0	46/82	56.1
10	36/56	64.3	36/65	55.4
11	46/67	68.7	46/71	64.8
12-14	118/146	80.8	118/151	78.1
15-17	55/64	85.9	55/65	84.6
18-20	24/25	96.0	24/25	96.0

TABLE 3
REMEDIAL COURSE PASS RATES VS. PLACEMENT TEST SCORES

Test Score	(No.Pass)/(No.Writing)	%	(No.Pass)/(No.Writing)	%
0-2	4/8	50.0	4/11	36.4
3-5	10/29	34.5	10/33	30.3
6	13/24	54.2	13/27	48.1
7	21/38	55.3	21/44	47.7
8	10/21	47.6	10/23	43.5
9	5/10	50.0	5/11	45.5
10	0/1	-	0/3	-
11	1/1	-	1/1	-
12-14	1/1	-	1/2	-
15-17	1/2	-	1/2	-

TABLE 4
STEPWISE DISCRIMINANT ANALYSIS RESULTS

Step Number	Question Entered	Subject Area	% Correct	Simplicity Rank
1	5	logarithms	50.2	10
2	8	trigonometry	67.9	4
3	3	algebra	54.6	9
4	17	word-geometry	57.4	7
5	6	logarithms	44.5	13
6	20	calculus	63.7	6
7	16	word-geometry	43.2	14
8	12	algebra	44.7	12
9	9	trigonometry	56.2	8
10	18	calculus	65.8	5
11	13	algebra	84.1	2
12	11	word-geometry	41.3	15
13	1	algebra	76.8	3
14	4	logarithms	33.7	4
15	10	functions	49.7	10
16	2	algebra	29.9	17
17	7	functions	92.2	1
18	14	word-geometry	7.4	20
19	15	calculus	18.3	18
20	19	calculus	8.3	19

TABLE 5
EFFECTIVENESS RATING OF PLACEMENT TEST QUESTIONS

Question	Chi-square	Effectiveness	Insertion Order
5	248.3	1	1
3	206.1	2	3
6	196.3	3	5
8	196.2	4	2
9	173.6	5	9
20	170.4	6	6
17	151.9	7	4
18	128.4	8	10
12	127.3	9	8
2	119.0	10	16
11	106.8	11	12
13	104.7	12	11
15	90.6	13	19
16	89.5	14	7
4	88.7	15	14
1	83.8	16	13
7	79.4	17	17
10	68.0	18	15
14	23.6	19	18
19	19.2	20	20

A GUIDE TO TEXTS ON THE DESIGN OF EXPERIMENTS

Gerald J. Hahn
Corporate Research and Development
General Electric Company
Schenectady, NY 12345

INTRODUCTION

There is now a large number of books on experimental design on the market. This paper provides an annotated bibliography of about 40 such texts, drawing in part upon the information given by the authors. (The abridged version included in these Transactions gives only the authors, titles and publishers of these books. The annotated bibliography can be obtained from the writer.) A summary table (included here) indicates the major topics covered, the technical level, the emphasized field of application and other pertinent information about each book. Comments on this guide are invited.

The determination of which books to include and exclude is necessarily a somewhat arbitrary one. In general, the guide is limited to books which deal principally with the design of experiments. Thus, books which are mainly concerned with the analysis of experimental data, and linear models have been excluded. Similarly, judgment was required in describing the books and in preparing the summary tabulation.

For an extensive bibliography on the literature of experimental design prior to 1968, see Ref. 1. This lists 8,378 works under 44 sub-headings, including one on bibliographies, which references about 150 additional works. A listing of the complete table of contents of 28 books on experiment and treatment design (most of which are included in the present bibliography) is also included.

Readers of this guide might also be interested in two handbook expositions on experimental design (Refs. 2 and 3).

REFERENCES

1. Federer, W.T. and Balaam, L.N., Bibliography on Experiment and Treatment Design Pre-1969, Hafner Publishing Company, New York, 1973.

2. Natrella, M.G., "Design and Analysis of Experiments," Chapter 27 in Quality Control Handbook, J.J. Juran, editor-in-chief, Third Edition, McGraw-Hill Book Company, New York, 1974.

3. Tingey, F.G., "Design of Experiments in Analytical Chemistry Investigation," Chapter 106 in Treatise on Analytical Chemistry, Part 1, I.M. Kolthoff and P.J. Elving, editors, Wiley-Interscience, New York, 1972.

LIST OF BOOKS

Anderson, V.L. and McLean, R.A., Design of Experiments - A Realistic Approach, Marcel Dekker, Inc., New York, 1974.

Bannerjee, K.S., Weighing Designs: For Chemistry, Medicine, Operations Research Statistics, Marcel Dekker, New York, 1978.

Box, G.E.P., Hunter, W.G., and Hunter, J.S., Statistics for Experimenters, John Wiley and Sons, New York, 1978.

Campbell, D.T. and Stanley, J.C., Experimental and Quasi-Experimental Designs for Research, Rand McNally and Company, Chicago, 1966.

Chakrabarti, M.C., Mathematics of Designing and Analysis of Experiments, Asis Publishing House, New York, 1962. (Not shown in 1978-1979 Books in Print)

Chapin, F.S., Experimental Designs in Sociological Research, Revised Edition, 1955, Reprinted in 1974 by Greenwood Press, Westport, CT.

Chew, Victor, ed., Experimental Designs in Industry, John Wiley and Sons, New York, 1958. (Not shown in 1978-1979 Books in Print)

Chilton, N.W., Design and Analysis in Dental and Oral Research, J.B. Lippincott Company, Philadelphia and Toronto, 1967.

Cochran, W.G. and Cox, G.M., Experimental Designs, 2nd ed., John Wiley and Sons, New York, 1957.

Cox, David R., Planning of Experiments, John Wiley and Sons, New York, 1958.

Daniel, C., Applications of Statistics to Industrial Experimentation, John Wiley and Sons, New York, 1976.

Davies, Owen L., et al., The Design and Analysis of Industrial Experiments, 2nd ed., Hafner Publishing Company, New York, 1956. (Reprinted by Longman, New York, 1978.)

Dayton, C.M., The Design of Educational Experiments, McGraw-Hill Book Company, New York, 1970.

Edwards, A.L., Experimental Design in Psy-

chological Research, 4th edition, Holt, Rinehart and Winston, Inc., New York, 1972.

Federer, Walter Theodore, Experimental Design: Theory and Application, Macmillan Company, New York 1955. (Not shown in 1978-1979 Books in Print.)

Federov, V.V., Theory of Optimal Experiments, Academic Press, New York, 1972 (Translated from the Russian by W.J. Studden and E.M. Klimko).

Finney, D.J., Experimental Design and Its Statistical Basis, The University of Chicago Press, Chicago, 1974.

Finney, D.J., An Introduction to the Theory of Experimental Design, University of Chicago Press, Chicago, 1975.

Fisher, Ronald A., Design of Experiments, 8th ed., Hafner Publishing Company, New York, 1966.

Gill, John L., Design and Analysis of Experiments in the Animal and Medical Sciences, 3 Volumes, Iowa State University Press, 1978.

Hersen, M. and Barlow, D.H., Single-Case Experimental Designs: Strategies for Studying Behavior Change, Pergamon Press, Elmsford, New York, 1975.

Hicks, Charles, R., Fundamental Concepts of Design of Experiments, Holt, Rinehart and Winston, Inc., New York, 2nd edn., 1973.

John, A., and Quenouille, M., Experiments: Design and Analysis, 2nd edition, Hafner Publishing Company, New York, 1977.

John, P.W., Statistical Design and Analysis of Experiments, Macmillan Publishing Company, Riverside, N.J., 1971.

Johnson, N.L. and Leone, F.C., Statistics and Experimental Design in Engineering and the Physical Sciences, Vol. 2, 2nd edn., John Wiley and Sons, New York, 1977.

Kempthorne, Oscar, Design and Analysis of Experiments, John Wiley and Sons, New York, 1952. (Reprinted by Krieger Publishing Company, Huntington, N.Y., 1975.)

Kimmel, H.D., Experimental Principles and Design in Pyschology, Ronald Press Company, New York, 1970

Kirk, R.E., Experimental Design: Procedures for the Behavioral Sciences, Brooks-Cole Publishing Company, Belmont, California, 1968.

LeClerg, E.L., Leonard, W.H. and Clark, A.G., Field Plot Technique, 2nd edn., Burgess Publishing Company, Minneapolis, MN, 1962. (Not shown in 1978-1979 Books in Print.)

Lee, W., Experimental Design and Analysis, W.H. Freeman and Company, San Francisco, CA, 1975.

Li, C.C., Introduction of Experimental Statistics, McGraw-Hill Book Company, New York, 1964.

Lindquist, E.F., Design and Analysis of Experiments in Psychology and Education, Houghton Miflin Company, London, 1956. (Not shown in 1978-1979 Books in Print.)

Lipson, C., and Sheth, N.J., Statistical Design and Analysis of Engineering Experiments, McGraw-Hill Book Company, New York, 1973.

Mendenhall, W., Introduction to Linear Models and the Design and Analysis of Experiments, Duxbury Press, Belmont, California, 1968.

Montgomery, D.C., Design and Analysis of Experiments, John Wiley and Sons, New York, 1976.

Myers, J.L., Fundamentals of Experimental Design, Allyn and Bacon, Inc., Boston, 2nd edition, 1972 (First edition, 1966).

Myers, R.H., Response Surface Methodology, Allyn and Bacon, Inc., Boston, MA, 1971. (Not shown in 1978-1979 Books in Print.)

Namboodiri, N.K., Carter, L.F., and Bialock, H.M., Jr., Applied Multivariate Analysis and Experimental Designs, McGraw-Hill Book Company, New York, 1975.

Ogawa, J., Statistical Theory of the Analysis of Experimental Designs, Marcel Dekker, Inc., New York, 1974.

Peng, K.C., The Design and Analysis of Scientific Experiments, Addison-Wesley Publishing Company, Reading, MA, 1967. (Not shown in 1978-1979 Books in Print.)

Raghavarao, D., Constructions and Combinatorial Problems in Design of Experiments, John Wiley & Sons, New York, 1971.

Vajda, S., The Mathematics of Experimental Design--Incomplete Block Designs and Latin Squares, Hafner Publishing Company, New York, 1967.

Winer, B.J., Statistical Principles in Experimental Design, Second Edition, McGraw-Hill Book Company, New York, 1971.

Yates, F., Experimental Design. Selected Papers of Frank Yates, Hafner Publishing Company, New York, 1970.

SUMMARY OF CHARACTERISTICS OF BOOKS

Characteristics	Anderson and McLean	Bannerjee	Box, Hunter and Hunter	Campbell and Stanley	Chakrabarti	Chapin	Chew	Chilton	Cochran and Cox	Cox	Daniel	Davies	Dayton	Edwards	Federer
Subject Coverage: Blocking designs	M-H	M-L	M	L	M	M	L-M	L	H	L-M	M-L	M-H	M	M	M-H
Factorial designs	M-H	M-L	M	L	M	M	M	M	H	L-M	H	M-H	M	M-H	M-H
Fractional factorial designs	M-H	M-L	L-M	M	L-M	M	M-H	M	M-H	L-M	H	M-H	L-M	L	L
Response - surface designs	M-H	M	M-H	M	M	M	M-H	M	M-H	L	L	M-H	M-L	M	M
Split plot situations	H	M	M	M	M	M	L	M	M-H	M	L-M	M	L	L-M	M-H
Nested situations	M-H	M	L	L	M	M	L	M	M	M	M	L-M	M	M	L-M
Number pages	418	141	653	84	120	297	268	365	611	308	294	656	441	488	591
Emphasis on calculations and analysis	L-M	L-M	L	M	M	L	L-M	M-H	L-M	M	L-M	L-M	M	M	M
References (approximate number)	150	83	150	150	37	100	350	200	150	100	100	50	50	125	340
Examples	M	L	M-H	M	M	M	M	M	M	M	M	M	M	M	L-M
Technical level	I	A	E	E	A	E	I	E	I	E	I	I	E	E-I	I
Applications emphasis	G/I	G	G/I	P	G	P	I	O	A/B	G	I	I	P	P	A/B

CODE

General: N: None; L: Little, M: Moderate, H: Heavy

Technical Level: E: Elementary (No background in statistics needed)
I: Intermediate (assumes one or two elementary statistics courses)
A: Advanced (assumes more than 1 or 2 statistics courses)

Applications Emphasis: A. Agriculture
B: Biology/Medicine
G: General
I: Industrial/Engineering/Scientific
M: Management/Business/Economics
P: Psychology/Education/Sociology/Behavioral Sciences/Social Sciences
O: Other

SUMMARY OF CHARACTERISTICS OF BOOKS

Characteristics	Federov	Finney (1974)	Finney (1975)	Fisher	Gill	Hansen and Barlow	Hicks	John and Quenouille	John	Johnson and Leone	Kempthorne	Kimmel	Kirk	LeClerg, Leonard and Clark	Lee
Subject Coverage: Blocking designs	N	L-M	M-H	M	M-H	N	M	M-H	M-H	M	M-H	M	M-H	M-H	L-M
Factorial designs	N	L	M-H	M	M-H	N	M-H	M-H	M-H	M	M-H	L-M	M-H	M	M
Fractional factorial designs	N	L	M-H	M	M	N	L-M	M	M-H	L-M	M-H	M	M-H	M	M
Response surface designs	N	N	L	N	M	N	L-M	M	M-H	M-H	N	N	N	N	N
Split plot situations	N	L	L-M	N	M	N	M	M-H	M	M	M-H	N	M-H	M-H	M-L
Nested situations	N	N	N	N	M-H	N	M	N	M	M-H	N	N	L	N	M-H
Number pages	292	169	223	245	409(1) 301(2) 182(3)	374	349	296	356	449	631	280	577	373	353
Emphasis on calculations and analysis	L	L	L	L	M-H	L	M	L-M	L-M	M-H	L-M	L	M	M	L-M
References (approximate number)	62	70	150	50	500	100	20	250	250	75	150	100	450	300	15
Examples	L	M	L	M	L-M	M	M	M	L-M	M	L-M	M	M	M	L-M
Technical level	A	E	I	E-I	I	E	I	I	A	I	A	E	I	E	I
Applications emphasis	G	B	A/B	A/B	A/B	P	I	A/B	G	I	A/B	P	P	A/B	P

CODE

General: N: None; L: Little, M: Moderate, H: Heavy

Technical Level: E: Elementary (No background in statistics needed)
I: Intermediate (assumes one or two elementary statistics courses)
A: Advanced (assumes more than 1 or 2 statistics courses)

Applications Emphasis: A. Agriculture
B: Biology/Medicine
G: General
I: Industrial/Engineering/Scientific
M: Management/Business/Economics
P: Psychology/Education/Sociology/Behavioral Sciences/Social Sciences
O: Other

SUMMARY OF CHARACTERISTICS OF BOOKS

Characteristics	Li	Lindquist	Lipson and Sheth	Menden-hall	Mont-gomery	J.L. Myers	R.H. Myers	Namboodiri, Carter and Blalock	Ogawa	Peng	Raghavarao	Vajda	Winer	Yates
Subject Coverage:														
Blocking designs	M	M	M	L-M	M	M	M	M	M-H	M	H	H	M	M-H
Factorial designs	M	M	L	L-M	M	M	L	M	M	M	M-H	M	M-H	M-H
Fractional factorial designs	M	M	L	L-M	M	M	L	L	M	M-H	M-H	M	L	M
Response - surface designs	M	M	M	L-M	M	M	H	M	M	M	M	M	M	L
Split plot situations	M	M	M	L	M	M	M	M	M	L-M	M	M	L-M	M
Nested situations	M	M	M	L-M	M	M	M	M	M	M	M	M	M	M
Number pages	460	393	518	464	418	465	246	688	465	252	386	120	907	296
Emphasis on calculations and analysis	M	M	M	L-M	M-H	M	M	M	M	M-H	M	M	M-H	L-M
References (approximate number)	35	50	100	100	85	100	50	250	220	90	500	109	100	100
Examples	M	L-M	M	M	L-M	L-M	L-M	L-M	L	L-M	L	L	M	M
Technical level	E-I	I	E	I	E-I	I	E-I	I	A	I	A	A	I	I
Applications emphasis	B/G	P	I	G/I	I	G	G	P	G	G	G	A	B/P	A

CODE

General: N: None; L: Little, M: Moderate, H: Heavy

Technical Level: E: Elementary (No background in statistics needed)
I: Intermediate (assumes one or two elementary statistics courses)
A: Advanced (assumes more than 1 or 2 statistics courses)

Applications Emphasis: A. Agriculture
B: Biology/Medicine
G: General
I: Industrial/Engineering/Scientific
M: Management/Business/Economics
P: Psychology/Education/Sociology/Behavioral Sciences/Social Sciences
O: Other

CASES: A SYSTEM OF CASE STUDIES FOR TEACHING STATISTICS

Charles H. Little, St. John's University
Michael P. Bennett, Xaverian High School

INTRODUCTION

A system of case studies is being developed to aid in the teaching of statistics. The system, called CASES, is completely computerized with the text of a problem, all instructions and computations available via a time-sharing computer system. The cases are actual business examples with the real world bugs remaining. The system emphasizes the interpretation of estimated statistics and de-emphasizes the mathematical formulas and computational techniques. The intent is to duplicate in the classroom an actual setting and to integrate instructions in statistics with use of an interactive computer time-sharing system.

A case study does not stand alone but must be supplemented with a statistics textbook and classroom lectures. Students are assigned a case, instructed how to do the computations on the terminal, and asked to come to class prepared to discuss the problem. A discussion is not totally structured but evolves according to students' questions and interests. Students can be given additional assignments to work outside singularly or in groups. The system is quite flexible and generates student interest in doing independent work on their own problems.

The system evolved around a required sequence of two statistics courses and a business forecasting course in the MBA program at St. John's University. The first course in introductory statistics includes descriptive statistics, probability distributions, sampling and testing. The intermediate statistics course includes bivariate linear regression, multivariate regression, curvilinear regression, analysis of variance, index numbers and time series analysis. The third course covers business forecasting techniques building on the tools from the previous courses.

When a student reaches the third course, he is prepared to do some rather extensive analyses on his own. The more complicated econometric and statistical techniques, such as the Box-Jenkins method, can be introduced. A detailed explanation of the technique is beyond the level of the course, but students are taught how to use a computer package and how to interprete the results. Through repeated use of different cases studies and discussions of the results, students gain a basic understanding of a technique and when to apply it. At a minimum, they learn enough to consult with an expert if they encounter problems when using a technique.

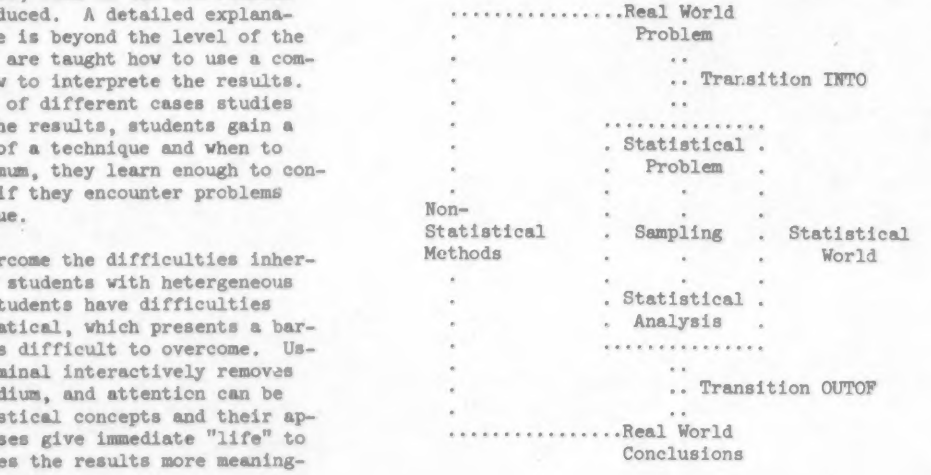
CASES helps overcome the difficulties inherent in a grouping of students with hetergeneous backgrounds. Many students have difficulties with anything mathematical, which presents a barrier for them that is difficult to overcome. Using the computer terminal interactively removes the computational tedium, and attention can be focused on the statistical concepts and their applicability. The cases give immediate "life" to the concepts and makes the results more meaningful by placing them in an actual setting. As a result, computational and mathematical blocks are forced to the background, if not overcome completely.

OBJECTIVES

The basic objective of CASES is to clarify the translation of a real world situation into a statistical problem and then the transition back to the real world with concrete recommendations for action. It is difficult for beginning practitioners of statistics to understand these two transitions. They often cannot see a connection between decisions to be made and a set of numbers to be analyzed. Quantification of a real world situation is difficult because many ideas and concepts cannot be strictly converted to numbers. The confusion is even worse in those situations where data are collected before a problem is formalized and then cannot be reconciled with the problem. The end result often is a tendency to concentrate on the numbers and to lose sight of the information they are conveying.

CASES forces a student to view problems as a whole. An analysis begins with a statement of the real world situation and then proceeds to a definition of the statistical population and the random variables. The student sees this transition with any accompanying compromises as an integral part of a whole analysis. Once the statistical computations are completed, statistical estimates and conclusions are converted to real world decisions in the context of the whole problem. Decisions are conditional on any compromises necessary at the start. A view of the whole situation eliminates the tendency to accept the numerical results blindly. Also, it prevents inferences so broad as to be useless in the real world. The process is shown in the chart below:

SEQUENCE OF A COMPLETE ANALYSIS



A complete analysis illustrates the advantage of using statistical tools. Data are viewed as information that can communicate, if played with properly and not restricted by rigid theoretical rules. A student sees that data can be consolidated, manipulated and massaged to obtain inferences. He sees the analytical tools and computational techniques only as the means to an end.

The case studies point up the strengths and weaknesses of various descriptive statistics and analytical tools. There is usually more than one way to analyze a problem, and the choice depends on a student's preferences. Procedural decisions are made as the analysis progresses, and the statistical inferences become cumulative. These inferences are seen as "good guesses" with accuracy estimates, not irrefutable "laws."

The system CASES offers considerable flexibility. For example, when at the computer terminal, a student may use canned routines, write his own supplemental programs or interface with computer packages. Flexibility is a particularly important element in a forecasting situation. Even though a model provides a good fit to past data, it very likely will require some modification when used to project into the future. Through interacting with the data and the analytical techniques, incremental changes are possible with the availability of more and/or different data.

REAL WORLD PROBLEM

The first step in the analysis is a concise statement of exactly what is known, what data are available and in what form. It is equally important to determine what is needed but not known. A comprehensive statement is essential at the start in order to establish the proper framework for analysis and to set bounds so that more is not expected than can be delivered with the available information. An analysis often is constrained by non-statistical considerations, namely a lack of time, money, data collecting and tabulating facilities. These limitations affect the scope of a problem and the final conclusions in the same as any statistical constraints.

The statement of the real world problem must include information on the limitations of the analytical methodology. Because of problems in quantifying concepts, it may not be possible to answer the real world questions directly, and compromises may necessitate a redefinition of the objectives. Compromises place bounds on a problem and conditions on any conclusions. It is important to have these spelled out at the start in order to come up with a realistic set of expectations. If expectations are too high, and end results may prove so disappointing that the whole study is tossed out, which amounts to "discarding the baby along with the bath water."

STATISTICAL PROBLEM

The transition to the statistical world is difficult, because the application of certain theoretical concepts is not always apparent. For example, many students find it difficult to define the appropriate statistical population-a task often difficult for the seasoned statistican. However, the population must include a definition of the elements of the population; both definitions depend on the objectives stated in the real world problem.

A determination of the appropriate observations or measurements on an element is the first step at quantification of the problem, and at this point the mathematics enters. Random variables and units of measurement must be specified according to the real world situation. Exact definitions may be complicated by measurement problems.

All the definitions must be consistent and produce a workable model, from which the nature of the random variables and the relations among the variables can be determined. The model establishes the type of analyses and the computational techniques to be used. At this point, assumptions are often necessary to fill in gaps and to make the model complete. It is important, however, that the problem not be assumed away. Students drilled in theoretical concepts tend to be too idealistic and want no loose ends. To tie up the loose ends, they wind up restricting the model to such an extent that no meaningful conclusions are possible.

SAMPLING

With the population firmly in mind, the next step is the selection of a representative sample of data, which is a crucial step in the analysis. Without a representative sample, inferences are not possible about the population and the study cannot continue. Often in an actual situation, data are collected before a study is designed, so the population is defined around the sample. This is true to a great extent in business research where considerable data come from secondary sources. The question then becomes can the data be used to answer the questions originally posed or does the focus of the problem have to be altered.

In the case studies, data are given along with as much detail as possible on how the sample was taken. A student must then decide if the sample is representative, and if not totally representative, what limitations are imposed as a result. Different sampling techniques are used for different case studies. The student is introduced to a sampling technique in the context of a real situation, and he sees exactly where sampling fits in the overall problem.

The key issue is how representative is the sample. The answer will greatly affect the transition back to the real world. It is usually not possible after an analysis to conclude that the sample was bad and the study should be repeated. Time and money constraints usually require some conclusions, regardless of how tentative they may be. If the sample is not totally representative, then the conclusions must be hedged until additional information is obtained.

STATISTICAL ANALYSIS

There are general guidelines and suggestions for doing the computations on the terminal, but a student is not restricted soley to these. He may try different alternatives as he goes along and is encouraged to combine the programs on the time-sharing system with his own routines. All data

are available via the terminal, but a student may enter additional data if he wishes. The student decides on appropriate critical values and tests as he proceeds. If problems are to be discussed in class, the information is given and not left to the student's choice, so that all the students will have common printouts for the discussion.

A student learns in the process how to effectively utilize a time-sharing terminal. The cases are not tied to a single technique or a single program. Other computing systems could be used, especially systems built around micro- and mini-computers. Data manipulation is emphasized, and a student is forced to use his own intuition and logic, which are the basic ingredients of all statistical analyses.

Statistical conclusions are necessary at each step of the computations as individual tests are examined. These results lead to additional tests and more conclusions. When results are inconclusive, supplemental information may be incorporated and different approaches tried. The interactive mode requires conclusions as the analysis progresses and these conclusions build upon each other.

An analysis eventually reaches the point where all the information must be put together for a complete solution to the problem. The task may prove to be very complicated, because it is possible that the results considered separately may seem contradictory. Differences must be reconciled, and this is made easier by following all the steps of the incremental analysis. There is no guarantee that the complete picture will provide all the answers; it is simply the best one given the available information.

REAL WORLD CONCLUSIONS

Finally, the most difficult step of the analysis looms, namely a return to reality. An accept or a reject the null hypothesis must be translated into a recommendation for action that could possibly involve millions of dollars. This is not a comfortable position for the statistician, which is the reason he likes to stop with a simple accept or reject. Usually that is not sufficient, and the statistican has to get involved in the real world. Students must be exposed to this phase of the analysis, because as practitioners, they must see that statistical decisions cannot be completely divorced from the consequences of the decisions.

The pieces of the overall statistical picture may not translate into a one to one correspondence with real world action. Care must be exercised not to extrapolate too far beyond the realm of the information analyzed. The transition is subjective and dependent on the expertise of the analyst. Different analysts could conceivably suggest different actions based on the same analyses. A student gets involved here by exploring the possibilities in a general discussion. The discussion is structured to allow a student freedom of choice, not to force him to come to a single, preconceived conclusion.

Original limitations, restrictive bounds, and conflicting statistics all add to a student's discomfort in having to come up with real world conclusions. He may be so overwhelmed that he decides no conclusions are possible. A favorite conclusion is that another, larger sample is needed. This is not acceptable, since it was decided earlier that the sample was representative of the population. A student must be forced to come up with some conclusions and recommendations or explain specifically why none are possible. He may finally suggest taking another sample and doing a different analysis, but he must include specific suggestions as how to proceed.

Students also have to be reminded that the numbers are only information, which are useful if they can be translated into real world decisions. For example, a 99 percent confidence interval with all its statistical accuracy may not survive the transition. It may be so wise that it is completely useless for planning purposes, and yet students are reluctant to discard the interval for whatever reason.

Finally, a student should go back and examine again the statement of the real world problem. He should reevaluate the initial shortcomings of the problem and then suggest ways to continue the analysis. Suggestions for continuation fare a logical end to any analysis whether or not it is physically possible to continue. The suggestions can be filed away for future use.

The greatest contribution of the system is the inclusion within the analytical framework of the transition INTO and the transition OUTOF the statistical world. The power and potential of statistical concepts come into focus when students learn to cope with the transitional problems. The analysis places the mathematical and statistical tools in their proper role and draws attention to the real world decision-making process.

The real world and statistical problems have to be considered together in order to achieve some degree of compatability. The transition INTO or the passage from actual expectations to precise statistical formulations cannot be standardized for every problem. Situations differ and each must be examined separately. Students must be made aware of this fact and given experience in doing just that.

The statistical analysis is approached in the context of data analysis, that is, manipulation of the data for maximum information. The object is to listen to the data, not twist it to fit a specific test or formula. Playing with data involves efficient use of computational techniques in order to minimize a student's time as well as computer time. The integration of computer instructions with the teaching of statistics becomes a must in this context.

The transition OUTOF or the passage to actual real world conclusions cannot be standardized either. Decisions must be reached logically for each situation and must not be restricted by an ill-fitting set of statistical theories. The students must gain confidence through experience to see the obvious, to admit errors and to search for reasons.

For the classroom discussions of the cases, students bring their own computer printout and

analytical results. They are all active participants in the discussions. Sometimes a consensus is reached in class on the results of the study; most often it is not. If not, differing results and conclusions are presented and discussed. A final decision rests with the individual student based on his own intutition, logic and understanding of the real world situation. The difficult part is to give the student the confidence to rely on himself to make a decision.

STATUS

CASES is still being developed, and there are approximately 15 case studies ready for use. Most of these have been used in class several times; the others are ready to be tried. It takes a couple of classroom discussions to assess students' reaction to the problem and to eliminate any instructional bugs.

The Rapidata time-sharing service is used for all work on the computer. In particular, the PROBE system is used for most of the computations. PROBE is a comprehensive language for the analysis and display of business data .

The system has been tested at the high school level, the college level and at seminars for businessmen with varying interests and backgrounds. It has proved very successful at all three levels. The results are promising enough that the system will be extended to cover additional statistical topics.

TRANSPARENCY MASTERS FOR CHARACTERISTICS OF PROBABILITY DISTRIBUTION AND THE NORMAL PROBABILITY DISTRIBUTION

Lawrence A. Sherr, University of Kansas
J. Earl Faulkner, Brigham Young University

The Audio-Visual Subcommittee of the American Statistical Association's Section on Statistical Education is in the process of preparing transparency masters for use by faculty in teaching the pre-calculus statistics course. The subcommittee envisions the publication of an inexpensive book of transparency masters from which an instructor can make his or her own transparencies.

It is planned the project will be composed of the following eight units:

1. Data, Frequency Distributions, and Probability Distributions
2. Characteristics of Probability Distributions (i.e., of random variables)
3. Special Probability Distribution - Binomial
4. Special Probability Distribution - Normal
5. Sampling (distributions)
6. Estimation
7. Tests of Hypothesis
8. Other Topics

The guidelines that have been used to develop these transparencies are:

1) the material is designed for the one-semester non-calculus introductory statistics course, though some of these transparencies will be useful for other courses,

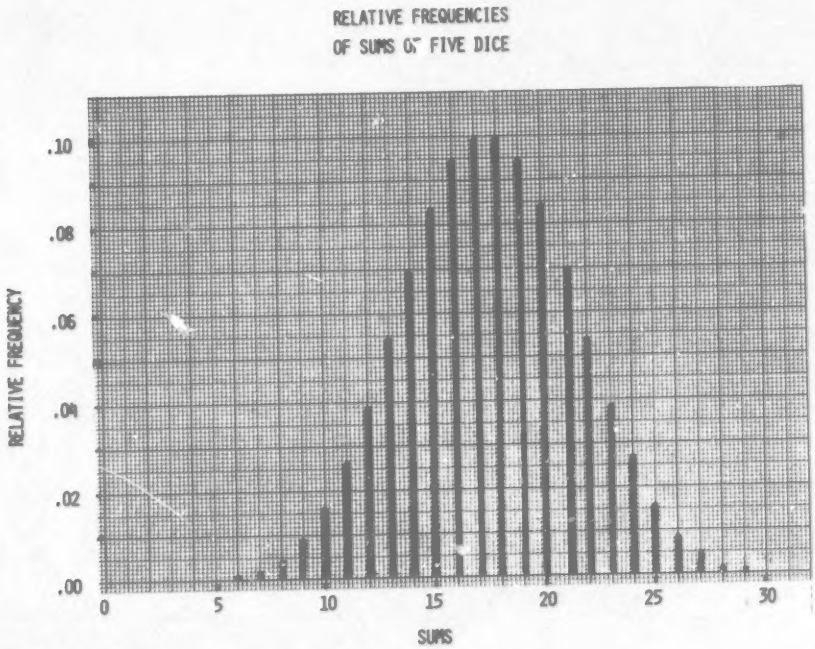
2. The transparencies should be for material we would like to put on the board but ordinarily do not because of the lack of time or drawing ability,

3. multi-overlay transparencies will be avoided to save time and money.

At this time four of the units are near completion.

The poster session presented both the transparency masters and examples of transparencies that can be produced from the masters for Units #2 and #4. In the exhibits following this discussion are five examples of the more than 30 masters that comprise these two units. From these masters it is both simple and inexpensive to make useful transparencies for teaching. Transparencies with black writing on a white background or colored background and with colored writing on a white background were exhibited.

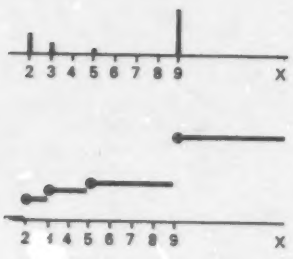
For further information please contact either of the authors or any member of the Audio-Visual Subcommittee.



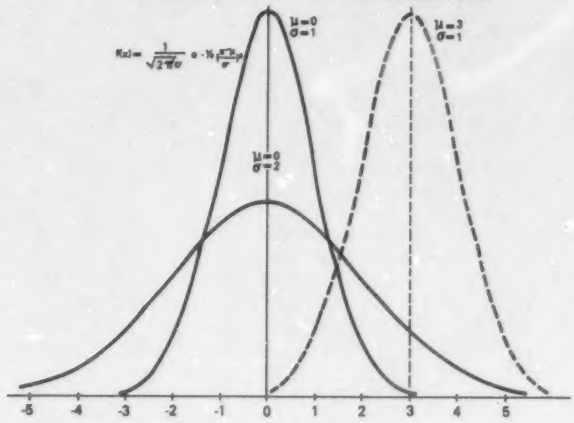
The Relationship Between a P.D.F. and a C.D.F.

X	2	3	5	9
f(x)	1/4	1/3	1/16	9/16

$$F(x)=\begin{cases}0 & X<2\\ 1/4 & 2\le X<3\\ 3/8 & 3\le X<5\\ 7/16 & 5\le X<9\\ 1 & 9\le X\end{cases}$$



The Effect of μ and σ in the Normal Distribution

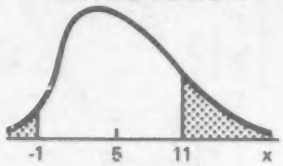


Chebyshev's Inequality:

Let X be a r. v. with mean μ and variance σ^2. Then for any $k>0$ we have:

$$P(|X-\mu|\ge k\sigma)\le\frac{1}{k^2}$$

Let X be a random variable with $\mu=5$ and $\sigma=3$
Find: $P(X\le -1 \text{ or } X\ge 11)$

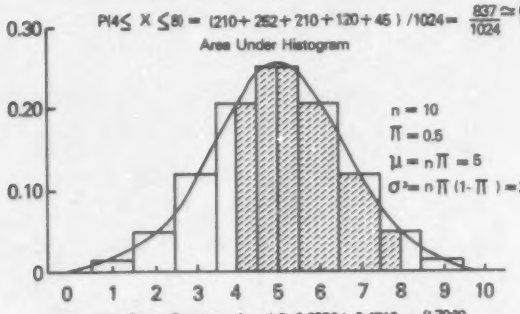


5−6	5	5+6
5−k(3)	5	5+k(3)
5−2(3)	5	5+2(3)

$$P(X\le -1 \text{ or } X\ge 11)=P(|X-5|\ge 2(3))\le 1/4$$

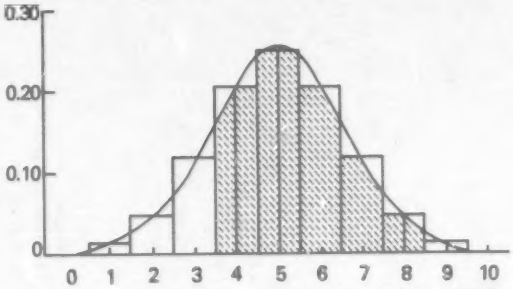
Normal Approximation to the Binomial

$P(4\le X\le 8)=(210+252+210+120+45)/1024=\frac{837}{1024}\simeq\underline{0.8174}$
Area Under Histogram



Area Under Curve Between 4 and 8: $0.2356+0.4713=\underline{0.7069}$

$Z_1=\frac{4-5}{\sqrt{2.5}}=0.63,\ Z_2=\frac{8-5}{\sqrt{2.5}}=1.90$ (No Adjustment for Continuity)



Area Under Curve Between 3.5 and 8.5 : $0.3289+0.4869=\underline{0.8153}$

$Z_1=\frac{3.5-5}{\sqrt{2.5}}=-0.95,\ Z_2=\frac{8.5-5}{\sqrt{2.5}}=2.21$ (Includes Adjustment for Continuity Represented by Shaded Areas)

WHAT'S IN A NUMBER? VERBAL EQUIVALENTS OF NUMERICAL p-VALUES

Robert F. Strahan and Paul G. McGovern

Iowa State University

"Galton's great gift lay in his awareness, which grew during his life, of the vagueness of many of the phrases in which men tried to express themselves in describing natural phenomena. He was before his time in his recognition that such vagueness could be removed, and a certain precision of thought attempted by finding quantitative definitions of concepts fit to take the place of such phrases (p.1). ...Though recognizable as a psychological condition of reluctance, or resistance to the acceptance of a proposition, the feeling induced by a test of significance has an objective basis in that the probability statement on which it is based is a fact communicable to, and verifiable by, other rational minds. The level of significance in such cases fulfils the conditions of a measure of the rational grounds for the disbelief it engenders (p.43)." (Fisher, 1960)

Still, and perhaps paradoxically, Fisher felt obliged--as do surely all instructors of basic statistics--to inform with verbal meaning the numerical p-values obtained. Thus, for example, in context of the χ^2 test of goodness of fit: "If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05, and consider that higher values of χ^2 indicate a real discrepancy." (Fisher, 1930)

To our knowledge, there has been no systematic investigation of the relation between numerical p-values and verbal expressions of belief/disbelief generally held by workers in statistics. This survey is a modest empirical effort toward bridging these two cultures of precise, but dumb, numbers and meaningful, but vague, words.

Method

The Survey Measure A questionnaire was constructed with these instructions:

Imagine that you conduct an experiment to compare the effectiveness of two new drugs. From a randomly drawn sample, half the subjects are assigned to the Drug A treatment group, half to the Drug B group. Measurements of drug effectiveness are obtained, and the means of the two groups are then compared by a t-test. Assume that all statistical assumptions have been met and that the experiment overall has been well performed.

The null hypothesis is that Drugs A and B are equally effective (their population means are the same); the alternative hypothesis is that the drugs differ in effectiveness. (Note that a 2-tailed test of significance is appropriate--if drug effectiveness is different, then either Drug A or Drug B is superior.

Below are verbal interpretations of a number of possible outcomes of this hypothetical experiment. In the blanks provided, please write your numerical estimates of the p-values that best correspond to these verbal interpretations. (A p-value, of course, is the probability, under the null hypothesis, of observing an outcome as deviant or more deviant than that actually observed.)

There then followed six statements of the form "I am _____ the drugs differ in effectiveness," with the underlined phrase in turn being "positive," "extremely sure," "very sure," "moderately sure," "slightly sure," and "undecided [whether]". (Details of the response format are given in the Results section.)

Subjects Copies of this questionnaire were given by mail to all staff and graduate students in the Statistics Department at Iowa State University, with a reminder notice sent one week later. Questionnaires were also administered to members of the first author's statistical methods course during a regular class period. These latter were predominantly graduate students in the behavioral sciences who had had at least one other statistics course.

Sixty-eight percent of the Statistics Department staff and 45 percent of the Statistics Department students who received questionnaires returned them. Twenty-two staff, 37 Statistics students, and 11 "Other" students provided usable protocols (i.e., questionnaires satisfying certain objective standards such as monotonicity of response)--70 subjects in all.

Results

One outlying observation (10^{-9} for the "positive" category) was winsorized (to 10^{-5}), and a logarithmic transformation yielded reasonably normal distributions of p-values. In Table 1 are found mean values (actually, antilogs of means of logs, rounded for clarity to one or two digits) for the three groups of subjects across the six verbal levels of sureness.

These values are very similar across samples, and a split-plot analysis of variance in fact showed no significant differences for subject groups or for the groups-by-p-value interaction--though of course the main effect for p-value was highly significant. (More piecemeal analysis--1-way Anovas for each of the six verbal labels followed by Tukey WSD comparisons--suggested only that the Statistics Department staff were more liberal --less demanding of objective evidence--in their use of the phrase "slightly sure ")

It was anticipated that certain "natural" or familiar p-values--.10, .05, .01, .00..--would be especially popular, and that was so. Table 2 shows a complementary and somewhat re-

Table 1

Mean p Values for Three Groups of Subjects

	Statistics Department Staff (n = 22)	Statistics Department Graduate Students (n = 37)	Other Students (n = 11)	
p ≤	.0002	.0004	.0002	I am positive--as positive as one can be about an empirical, statistical question--the drugs differ in effectiveness.
p =	.002	.002	.002	I am extremely sure the drugs differ in effectiveness.
p =	.01	.01	.01	I am very sure the drugs differ in effectiveness.
p =	.07	.05	.04	I am moderately sure the drugs differ in effectiveness.
p =	.15	.09	.08	I am slightly sure the drugs differ in effectiveness.
p ≥	.20	.13	.18	I am undecided--really can't judge one way or the other--whether the drugs differ in effectiveness.

Table 2

Frequencies of Verbal Category Endorsements for Selected p Values

	p = .001	p = .01	p = .05	p = .10
Positive	20	7	1	
Extremely Sure	29	15	0	
Very Sure	4	34	14	
Moderately Sure		7	34	17
Slightly Sure			11	23
Undecided			5	12

dundant picture of the results. The data here are of course a (large) subset of the total observations.

Discussion

The first of the tables shows a considerable degree of uniformity, on average, in the coordination of verbal and numerical indices of surety. It is interesting that this uniformity obtains not just for three different samples, but as well for three samples whose members differ rather widely in extent of statistical expertise/experience.

The second table shows the degree of variability in choice of verbal labels attached to several commonly used p-values (as might have been expected, this variability was more pronounced in the intermediate p range). Table 2 also permits tentative assignment of "consensual" or "average" verbal meaning to three traditional p's: for .10, "slightly sure;" for .05, "moderately sure;" for .01, "very sure"--at least; and for .001, "extremely sure"--at the very least.

Strictly speaking, of course, these results hold only for the particular instrument employed. We wanted a statistical problem of intermediate detail. We wanted a context sufficiently specific that the issue posed would be readily grasped and reasonably realistic, but one not so specific that extraneous--for our purposes--factors such as prior probabilities, utilities, risk functions, the a priori falsity of all point hypotheses, etc., should become salient. Our choice was obviously a compromise.

Similarly, we do not claim that our selection of verbal categories ("undecided" to "positive") was ideal--in particular, the expression "slightly sure" may have fallen awkwardly on some ears and minds, and even with its empirical qualification "positive" was offensive to some. Still, this selection was based on previous scaling studies of diverse substantive content, and superior alternatives are not conspicuous.

Fisher Once More The question arises: How would Fisher have responded to this questionnaire? One hesitates to put numbers in his mouth; still, when the cat's away, some mice will play. Our speculation:

Positive	.01
Extremely sure	.01
Very sure	.01
Moderately sure	.05
Slightly sure	.10
Undecided	.10

The bases for this easy arrogance are Fisher's writing in the two previously cited books and in Statistical methods and scientific inference (1959), from all three of which excerpts follow.

Of $p > .10$: "Certainly no reason to suspect the [null] hypothesis."

$p < .05$: "Convenient...as a standard level of significance...." "Convenient ...as a limit in judging whether a deviation is to be considered significant or not." "We shall not often be astray if we draw a conventional line at .05, and consider that higher values...indicate a real discrepancy." "The observed value...just exceeds the 5 per cent point, and the experimental result may be judged significant, though barely so....The example chosen, in fact, falls just on the borderline between those results which can suffice by themselves to establish the point at issue, and those which are of little value except in so far as they confirm or are confirmed by other experiments of a like nature."

$p < .02$: "Strongly indicated."

$p < .01$: "Clearly significant." "Very significant."

$p < .000001$: "Provided the deviation is clearly significant, it is of no practical importance whether P is .01 or .000001."

Assuming some truth to our speculation, one might wonder why Fisher would have accepted less compelling evidence than the actual subjects of this research in order to be "extremely sure" or "positive" that the two drugs did in fact differ. The assumed discrepancy may simply be artifactitious. Had Fisher actually responded to the questionnaire, rather than his responses having been inferred from his writings, there might have been no discrepancy; the obviously ordered verbal categories do seem to invite a spreading out of p-values (an invitation nevertheless spurned by several actual respondents).

A more substantive explanation may be the embarrassment of riches the computer has given us: we are quite used, as Fisher was not, to considering hundreds, sometimes thousands, of statistical outcomes, and consequently are not so reluctant as Fisher might have been to think that a "rare and unusual" event has happened to us (still less, perhaps, to someone else). Despite the fact that the present "study" was presented as a single, planned experiment, this erosion of the significance of statistical significance may nonetheless have influenced the real subjects' responses in the conservative direction.

In this connection, and with obvious relevance to the previously mentioned possibility of method artifact, one might wonder what the results of this investigation would have been had the problem been posed along these more personal lines: "Imagine that you are given a coin and asked to flip it a number of times in order

to determine whether it is fair." We suspect that most persons would not require an outcome so dramatic as "less than two chances in a thousand" in order to be "extremely sure" that the coin was not fair. But we may be wrong.

In any event, speculation on what Fisher might or might not have said is mainly of historical interest. What this survey does seem to provide for current and future usefulness is evidence that, notwithstanding a certain degree of individual differences, there is nevertheless a rather impressive consensus with regard to the correspondence between verbal and numerical indices of belief/disbelief in a statistical hypothesis.

We are at least moderately sure of this conclusion, and are certainly positive that we are not undecided. We think.

References

Fisher, R. A. Statistical methods for research workers. London: Oliver & Boyd, 1930.

Fisher, R. A. Statistical methods and scientific inference. New York: Hafner, 1959.

Fisher, R. A. Design of experiments. New York: Hafner, 1960.

Acknowledgement

We thank, of course, the many colleagues and students who participated in this study.